Defining the protein complement of CpG islands

John Paterson Thomson

Thesis presented for the degree of Doctor of Philosophy

The University of Edinburgh

2011

~~ In memory of my role model and father, John P Thomson. ~~

Declaration

I declare that this thesis was composed by myself and the research presented is my own unless otherwise stated.

John Paterson Thomson

2011

Acknowledgements

I would firstly like to thank Professor Adrian Bird for taking a chance on me and for giving me the opportunity to carry out my PhD research at the University of Edinburgh. Without his input and assistance I would have struggled to complete this work. Many thanks must also be given go to Cancer Research UK for the funding they provided during my PhD. To Christine Struthers I owe many thanks not only for looking after the administrative side of things on my behalf but for emailing my bank details out to the lab during my first week! Luckily I have no money anyway.

For all their help both professionally and personally I wish to thank Aileen Greig, Jim Selfridge and John Connelly. Without your advice and support I would have never have made it this far so I most definitely owe you a drink (or ten!). Special thanks go out to Shaun Webb and Alastair Kerr for all their help and assistance with the bioinformatics aspect of my project. I must also acknowledge everyone who has contributed to this work by giving their advice, their reagents or by proofreading this thesis, specifically to Pete Skene, Robert Illingworth, Heather Owen and Robert Ekiert, whose advice and assistance made this work possible.

Finally I wish to thank all of the people outside of the lab who have supported me through my PhD project and throughout my life in general. Firstly to my beautiful partner Gillian for all of the support she has given me throughout this work and for giving me two such beautiful children. And to my family, specifically my parents; you were there for me thought the years to keep me going and I owe any success I achieve to you. Although we are facing tough and challenging times I know we will pull through.

Abstract

In higher eukaryotes, the DNA base Cytosine can exist in a variety of modified forms when in the dinucleotide CpG. Although a methylated form tends to dominate within the genome, approximately 1% of all CpG dinucleotides are found unmodified at high densities spanning around 1Kb and tend to co-localise to the 5' ends of around 60% of annotated gene promoters. These unique DNA sequences are known as CpG islands (CGIs) and their role within the genome to date is largely unknown. Methylation of CGIs in cancers however has been linked to silencing of associated genes implying a role in gene regulation. Furthermore these sites are also interesting as they remain specifically non-modified within a genome rich in methylated CpG.

We set out to better understand the roles for CGIs through the characterisation of any specific CGI binding proteins. Digestion of nuclei with methyl sensitive restriction enzymes facilitates the purification of CGI fragments. Subsequent immunohistochemistry on the CGI chromatin fragments along with ChIP-PCR over several CGIs revealed an enrichment of the "active" histone modifications including H3K4me3, a depletion of the "silencing" marks such as H3K27me3, as well as a group of CGI specific binding factors. These latter proteins contained a domain previously shown to bind to non-methylated CpG dinucleotides (the CXXC domain) and as such were ideal candidates for CGI specific factors, in particular a protein called Cfp1. Genome wide sequencing revealed a striking correlation between Cfp1 and H3K4me3 which were both seen at around 80% of islands. Furthermore, the presence of Cfp1/H3K4me3 at islands tended to have a negative correlation with the presence of chromatin rich in the silencing histone modification H3K27me3. Closer investigation of the Cfp1 protein reveals it to be a true non-methyl CGI binding factor *in vivo* and shRNA reduction of Cfp1 levels to around 10% of wild type resulted in a precipitous drop in H3K4me3 levels over CGIs without a dramatic reduction in global H3K4me3 levels. As Cfp1 has been shown to be part of the Set1 histone H3K4 methyltransferase complex responsible for this modification, this CXXC protein may be attracting this histone modifying complex and as such represents a method whereby the underlying DNA sequence (CpG) can drive the overlying epigenetic state. This study may go some way to understanding the functional significance of CGIs within the genome.

TABLE OF CONTENTS

Figures

Tables

Abbreviations

5AzaC	5-Azacytidine
5hmC	5-hydroxy methylcytosine
5meC	5-Methylcytosine
Ac	Acetyl
Actb	Actin-β
Ape1	AP endonuclease
Ash1	Absent Small or Homeotic discs 1
Atg9b	Autophagy related 9 homolog B
BDNF	Brain derived neurotrophic factor
BPTF	Bromodomain PHD finger transcription factor
C5	Carbon-5
CAF	chromatin assembly factor
CAP	CXXC affinity Purification
Cfp1	CpG binding Protein
CGI	CpG island
CTD	carboxy-terminal domain
Dnmt1/2/3a/3b/3L	DNA methyltransferase 1/2/3a/3b/3L
ECL	Enhanced chemiluminescence
ENCODE	Encyclopedia Of DNA Elements
ES Cell	Embryonic Stem Cell
Esc/Eed	Extra sex combs
EzH2	Enhancer of Zeste
E_2-ERα	Estradiol-estrogen receptor
GEO	Gene Expression Omnibus
GNAT	Gcn5 related acetyltransferase
HAT	histone acetyltransferase
HDAC	histone deacetylase
HMTase	Histone methyltransferase
HTFs	HpaII Tiny fragments
H3K4me	Histone H3 lysine 4 methylation
H3K4me1/2/3	Histone H3 lysine 4 mono-/di-/tri- methylation

H3K9me1/2/3	Histone H3 lysine 9 mono-/di-/tri- methylation
H3K27me1/2/3	Histone H3 lysine 27 mono-/di-/tri- methylation
H3K36me1/2/3	Histone H3 lysine 36 mono-/di-/tri- methylation
hmCpG	Hydroxymethyl CpG
hmC	Hydroxymethyl Cytosine
IGB	Integrated Genome Browser
ING	Inhibitor of growth
JmjC	Jumonji C domain proteins
Kdm2a	Lysine demethylase 2a
MBD	Methyl binding domain
meCpG	Methyl CpG
MeCP2	Methyl-CpG binding protein
MLL1/2	mixed-lineage leukaemia 1/2
ncRNA	non-coding RNA
NuRD	Nucleosomal remodelling & deactylase complex
NURF	Nucleosome remodelling factor
NLS	Nuclear localisation signal
nts	nucleotides
Oprd1	Opiod receptor delta 1
ORC	Origin replication complex
P-M-	*P53-/- Dnmt1 -/-* Mouse embryonic fibroblast cell line
P53-	*P53-/-* Mouse embryonic fibroblast cell line
Pc	Polycomb
Ph	Polyhomeotic
PRC1	Polycomb repressor complex 1
PRC2	Polycomb repressor complex 2
Psc	Posterior Sex combs
qPCR	Quantitative polymerase chain reaction
RNAP II	RNA Polymerase II
RISC	RNA-induced silencing complex
Sce	Sex combs extra
siRNA	Short interfering RNA

Suvar3-9	Suppressor of variegation 3-9
Su(z)12	Suppressor of Zeste-12
Taf1	TATA box protein associated factor
Tet1	Ten-eleven translocation-1
TGD	thymine DNA glycosylase
TRD	transcription repression domain
TRDMT1	tRNA aspartic acid methyltransferase 1
TRX	Trithorax
TSS	Transcriptional start site
Xi	Inactive X chromosome

Chapter 1: Introduction

In biology, the term epigenetics is generally described as the phenomenon whereby the overall phenotype and gene expression profiles of a cell or organism are controlled by mechanisms other than the underlying DNA sequence. This name, derived from the Greek "epi" meaning over or above, describes the observation that the DNA sequence is under a higher level of control outside of its sequence. This control is brought about both by the direct chemical modification of the DNA itself along with the regulation of DNA binding proteins known as the histones. The wrapping of DNA around these histone proteins form an octameric protein structure called the nucleosome and the subsequent compaction of these structures packages the DNA into the nucleus of a cell. Therefore to gain access or to attract and repel certain enzymatic complexes to the DNA these nucleosomes must be regulated tightly.

Epigenetic control is most prominent during development as pluripotent embryonic stem cells (ES cells) containing the same DNA sequences, differentiate into many functionally diverse cell types. These differentiated cells contain the same DNA sequence as the ES cells, barring any mutations picked up during differentiation; however the gene expression profiles of these cell types will vary massively. Many changes in gene expression are brought about through epigenetic mechanisms such as the modification of the local chromatin structure (through modulation of nucleosome positioning or modification of histone tails) or through epigenetic modification of the DNA without changes to the sequence itself. In turn, these modifications can bring about changes in gene expression through the attraction or exclusion of specific binding factors which can facilitate or repress the transcriptional machinery. As the field of epigenetic research has grown, the importance of such control mechanisms has become apparent in a range of developmental disorders and diseases (Egger, Liang et al. 2004). These include many of the autism spectrum disorders (Schanen 2006) as well as carcinogenesis and the progression to tumourogenesis (Jones 2002). Furthermore several disorders have been found to contain "epimutations" which occur without any detectable changes or mutations in the DNA sequence itself (Horsthemke 2007). Investigating the links between specific histone modifications,

chemical modifications to the DNA sequence and the proteins recruited to such modified sites should lead to a better understanding of epigenetic gene control.

1.1 Chromatin structure and post translational modification of the nucleosome

The basic repeating structure of chromatin is that of the nucleosome; consisting of ~146 base pairs of DNA wrapped around a octamer of the four histone proteins H2a, H2b, H3 and H4[i]. This octamer is made up of two dimers of H3 and H4 and two dimers of H2a and H2b (each forming tetramers). These two tetramers then combine with ~146bp of DNA to form the octameric nucleosomal structure (fig 1.1-1, A; right). This nucleosomal unit is responsible for the compaction of the DNA sequence of an organism into chromosomes through the folding of these DNA:protein complexes into higher order structures (fig1.1-1, A; left). Not only does the nucleosome facilitate the packaging of the DNA, it also plays a critical role in the regulation of the genetic code by restricting the access of sequences within these compact structures. This is thought to primarily occur through the modification of the histone proteins which can affect the compaction of local nucleosomes. The histone proteins consist of both a globular domain and an N terminal flexible tail. Whislt the globular domains interact with one another to form the octameric histone core complex, the tails protrude out from this structure and act as platforms for chemical modification by a host of enzymatic complexes (Figure 1.1-1, B). Chemical modification of certain residues in these tails results in changes to both the overall compaction of the chromatin as well as attracting or repelling specific protein complexes. Thus the higher order fibers can be decompacted (forming a “euchromatic” environment) allowing proteins to access and exert an effect on the DNA, or tightly compacted (forming a “heterochromatic” environment) resulting in inaccessible DNA sequences. This form of regulation when placed upon the DNA is regarded as epigenetic as it can affect the regulation of gene expression outwith the DNA sequence.

i. Although typical nucleosomes consist of the histones mentioned in the text, nucleosomes are found which contain unique histones variations such as the H2a variant H2a.z and the H3 variant CENP-A.

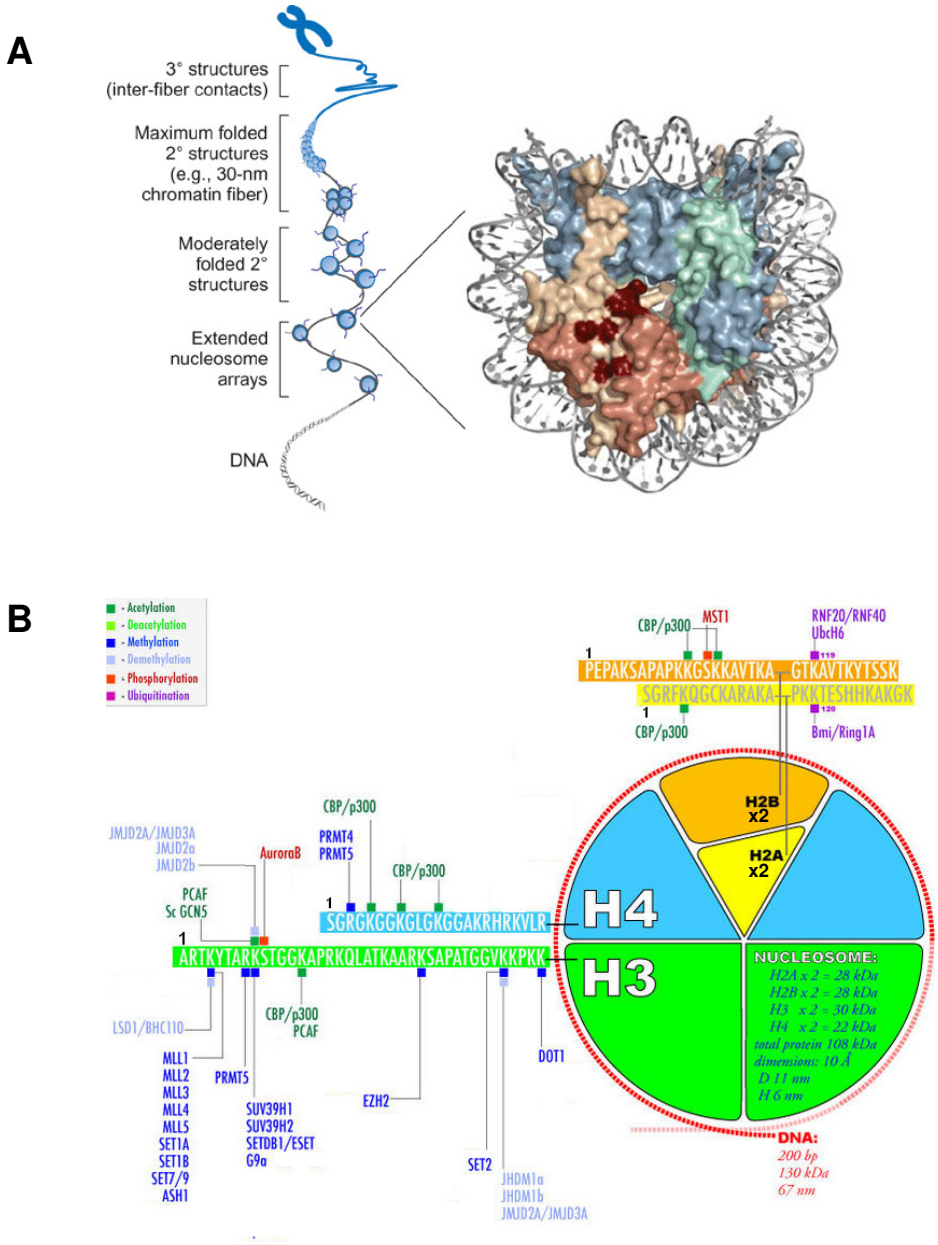

Figure 1.1-1. The structure and modification of the mammalian nucleosome. **A**.The structure of a typical nucleosome is shown on the right. Histones H2A (yellow), H2B (light red), H3 (green) and H4 (blue). Residues comprising the charged pocket are shown in dark red. Not shown are the flexible N terminal histone tails which are freely associable outside the nucleosome. These nucleosomes are packaged into higher order structures as shown on the left, which finally results in the formation of the chromosome.**B.** Schematic of selected histone tail modifications along with the organisation of the nucleosome. Some of the key post translational modifications are shown across the four histone tails (see key for colour coding).

Fig A from Catereno & Hayes, 2007. Fig B courtesy of R.Meehan at the MRC, Edinburgh

Study of these histone tail modifications and their effector proteins as well as the distribution of these marks throughout the genomes of multiple organisms has resulted in a greater understanding of epigenetic regulation. Nevertheless many of these findings are purely correlative and as such require further investigation. The most commonly investigated histone modifications are the incorporation of methyl or acetyl groups into lysine and arginine residues or the phosphorylation of serine residues across the four histone tails. The specific modifications placed upon certain residues in turn exert a particular effect upon the regulation of genes.

1.1.1 Selected histone tail modifications

The N terminal tails of the four histone proteins can be post translationally modified at many residues (Figure 1.1-1). Some of these modifications have been studied more extensively than the others; largely due to the roles such modifications appear to play in the regulation of several key processes such as gene regulation, cell cycle progression and overall chromatin compaction. Some of the more frequently researched modifications are explained below in greater detail.

i) Methylation at lysine 4 of Histone H3

Methylation of the fourth lysine residue on the N terminal tail of Histone 3 (H3K4me) was first identified in trout testes in 1975 (Honda, Dixon et al. 1975) and is found to exist in three states; mono-, di- and trimethylated H3K4. This methylation is progressive from the mono-methylated state (H3K4me1) through to a fully tri-methylated form (H3K4me3) which is generally found to be associated with transcriptionally active chromatin (Bernstein, Kamal et al. 2005). Studies in yeast have shown that the mono- and di-methyl marks occur at both active and inactive genes (Santos-Rosa, Schneider et al. 2002) often peaking in the body of genes (Pokholok, Harbison et al. 2005). Recent high resolution sequencing studies in the human genome revealed a somewhat contrasting pattern for the H3K4 mono- and di methylation in vertebrates (Barski, Cuddapah et al. 2007). Here it was found that although the mono- and di- methyl marks positively correlate with gene transcription they tend to peak at around -900 and +1000bp for mono H3K4me1 and -500 and +700bp for di-methyl over relative to the transcriptional start site (TSS) of genes (Figure 1.1-2) (Barski, Cuddapah et al. 2007). The trimethylated form of the modification is known to associate specifically with genes which are

active or destined to be so (Bannister, Schneider et al. 2002; Bernstein, Humphrey et al. 2002; Santos-Rosa, Schneider et al. 2002; Schubeler, MacAlpine et al. 2004). In budding yeast the H3K4me3 mark is seen at the 5' end of genes during transcription and is thought to be set up through a combination of factors including RNA polymerase II (RNAP II) and a histone methyltransferase (HMTase) complex called COMPASS (Miller, Krogan et al. 2001; Ng, Robert et al. 2003). Studies on mammalian cells reveal a tight correlation between the TSS and H3K4me3 modified histones, with this modification found enriched from -300 to +100bp over promoters (Figure 1.1-2) (Barski, Cuddapah et al. 2007).

This modification may provide a platform for the binding of chromatin modifying complexes which contain a PHD domain specific for methylated H3K4 residues. One such protein is the "bromodomain PHD finger transcription factor" (BPTF) which is part of the nucleosome remodelling factor (NURF) complex (Mizuguchi, Tsukiyama et al. 1997). Attraction of BPTF results in the formation of an accessible euchromatic state, allowing the binding of the transcriptional machinery to the DNA sequence. The H3K4me3 mark can also attract the Inhibitor of growth (ING) protein via its PHD domain. This protein has been shown to bind to both a histone actetyltransferase (HAT) complex as well a histone deacetylase (HDAC) complex, providing a potential link between histone acetylation and H3K4 methylation (Russell, Berardi et al. 2006). Finally, the histone H3 lysine 9 (H3K9) demethylase Kdm3a has been shown to bind to the H3K4me3 modification and removes any methylation from H3K9 residues, a mark normally associated with silencing. This results in the perpetuation of an active chromatin state, not only through the attraction of activating complexes, but through the loss of attraction of repressive complexes (Yamane, Toumazou et al. 2006). As studies into these H3K4me3 binding proteins are limited much is still left to learn regarding the downstream effects of this histone modification.

ii) Methylation of lysine 27 on histone H3

Methylation of lysine27 on histone H3 is a repressive histone modification associated with heterochromatic regions of the genome as well as at the promoters of silenced genes (Boyer, Plath et al. 2006; Roh, Cuddapah et al. 2006; Barski, Cuddapah et al. 2007). To date the trimethylated from of this modification is the best understood and histones rich in H3K27me3 tend to be found over much

larger domains (averaging 5-20kb) than the discrete peaks averaging 3kb seen for H3K4me3 (Figure 1.1-2) (Bernstein, Mikkelsen et al. 2006). Recent work has also revealed the presence of "bivalent" regions of chromatin containing both H3K4me3 and H2K27me3 marks, most frequently found at a proportion of promoters in pluripotent ES cells (Bernstein, Mikkelsen et al. 2006). These bivalent domains are thought to represent regions "poised" for later transcription. It appears that during differentiation either the H3K4me3 or H2K27me3 modification is lost from such sites depending on the expression of the gene in the differentiated cell type.

The H3K27me3 modification is bound by a chromodomain containing protein termed Polycomb (Pc) which is part of a larger complex of proteins called the polycomb repressive complex 1 (PRC1, (Messmer, Franke et al. 1992). This complex contains the aforementioned polycomb protein along with stoichiometric amounts of three other proteins; Posterior sex combs (Psc), Polyhomeotic (Ph)

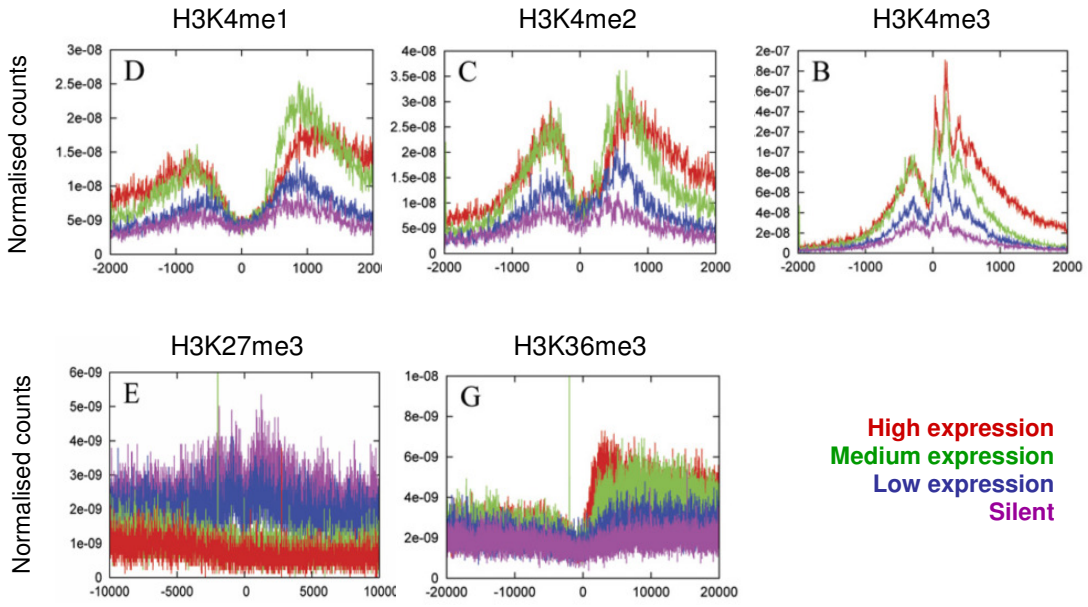

Figure 1.1-2. The patterns of selected histone methylations over human TSS. ChIP sequencing across the human genome generating high resolution plots for H3K4me1, 2 & 3, H3K27me3 and H3K36me3 modifications. The plots are grouped based on the relative activity of the associated gene (see key, bottom right). Each signal is plotted over a region spanning the TSS ("0"). H3K27me3 and H3K36me3 plots are plotted over a larger region than the H3K4 methylations as these modifications tend not to form discrete peaks. Signals are displayed as normalised counts revealing peaks and troughs for each of the histone modifications. For more on ChIP-sequencing see chapter 4.4. Figures from (Barski, Cuddapah et al. 2007)

and Sex combs extra (Sce) (reviewed by (Gil and Peters 2006). Although this complex is relatively simple in Drosophila, multiple orthologues of these proteins exist in mammals (Figure 1.1-4 below). As PRC1 binding occurs over large regions of H3K27me3 the end result is a large region of inaccessible and compacted chromatin which provide stable, long term and heritable silencing of underlying genes.

iii) Methylation of lysine 36 on histone H3

The methylation of histone tails at lysine 36 (H3K36me3) is mechanistically linked to the progression of the transcriptional machinery through the gene body. This mark is generally enriched within the coding regions of genes whilst low or absent over the 5' promoter regions (Figure 1.1-2) (Barski, Cuddapah et al. 2007; Blackledge, Zhou et al. 2010). The functional significance of this modification is thought to be mediated through the attraction of specific effector proteins. During transcription the RNAP II complex passes through the gene body in a process termed transcriptional elongation. As the body of genes are thought to be found in a compacted chromatin state, the progression of the RNAP II complex relies on the localised de-compaction of this chromatin environment. This is brought about through the acetylation of histone tail residues (see section 1.1.1 iv below). In order to prevent spurious transcription, the state of the chromatin must be tightly regulated once the RNAP II complex has passed. This regulation is thought to be modulated through the H3K36me3 modification, which attracts the EAF3 chromodomain containing protein (Joshi and Struhl 2005). This H3K36me3 binding factor in turn attracts the histone deactylase complex Rpd3S which removes the acetyl groups from the histones once the transcriptional machinery has passed. The removal of acetyl groups from the histone tails results in a "closed" chromatin state preventing the initiation of internal transcripts. As such, the H3K36me3 modification is often used as a method to identify actively transcribed genes. Promoters however, are strictly devoid in this histone modification, presumably as they are required to become acetylated during transcriptional initiation.

iv) Acetylation of residues on the H3 tail

The addition of acetyl groups to histone tails can occur at a large number of the arginine and lysine residues on both histones H3 and H4 (Figure 1.1-1). These modifications are tightly associated with

transcriptional activity and are found at the promoters of actively transcribed genes (Figure 1.1-2) as well as associated with the elongating RNAP II complex during transcription (Liang, Lin et al. 2004; Schubeler, MacAlpine et al. 2004). The mechanisms by which this modification brings about a decondensed chromatin state is not known however several models have been proposed. A likely mechanism is that the neutralisation of positively charged lysine residues in the histone tails by the addition of acetyl group reduces the interaction between the histone and the negatively charged DNA. Alternatively, the attraction of chromatin modifying complexes containing a bromodomain specific for the acetyl mark may result in a change in the local chromatin state. Evidence exists suggesting that the chromatin remodelling complex Swi/Snf, is bound to acetylated H3K23 residues resulting in nucleosomal remodelling (Awad and Hassan 2008). Additionally the TATA box protein associated factor (Taf1), which is part of the RNA polymerase II complex, is seen to bind to acetylated H4 tails through its bromodomain, resulting in a recruitment of the complex to modified histone tails (Matangkasombut and Buratowski 2003).

1.1.2 Regulation of the histone tail modifications

As the post translational modifications placed on the histone tails give rise to important regulatory mechanisms, these marks must be tightly regulated. As such, each modification is regulated to ensure the coordinated addition and removal in a temporal fashion. Some of these modifications are relatively simple in their regulation whilst for others several multiprotein complexes act.

i) Regulation of H3K4 methylation by the Trithorax group of proteins

The enzymes responsible for the methylation of histone H3K4 were first identified in *Drosophila* as having essential roles in the maintenance of gene expression (Byrd and Shearn 2003). This protein, known as “Trithorax” (TRX) and “Absent Small or Homeotic discs 1” (Ash1) contains a catalytic SET methyltransferase domain responsible for the methylation at H3K4 residues. Enzymatic complexes containing this conserved SET domain were also found in the yeast and mammalian genomes indicating that similar methylation role existed for these proteins (Miller, Krogan et al. 2001). The yeast homologue of trithorax, Set1, is part of a large multiprotein complex called COMPASS. Characterisation of this 400kd complex revealed six other proteins (Figure 1.1-3),

CPS60, 50,40,35,30 and *25* (Miller, Krogan et al. 2001; Shilatifard 2008) in addition to the catalytic Set1 protein. Further work on this complex found that the Set1 protein is able to bind to the phosphorylated carboxy-terminal domain (CTD) of RNA Pol II and catalyse the methylation of H3K4 residues (Ng, Robert et al. 2003). Of the other proteins, *CPS50*, *35* and *30* contain WD40 repeats (Figure 1.1-3). WD40 domains are around 40 amino acids long and are thought to be important within the formation of multi-protein complexes (Li and Roberts 2001). Other interesting proteins include *CPS60* (also known as *Bre2*) and *CPS40*, also known as *Spp1*. These two proteins have been shown to be important for the progression of the methylated lysine from a dimethylated form to a trimethylated form (Schneider, Wood et al. 2005). Furthermore as *Spp1* contains two PHD domains it is thought to be important in targeting the complex to chromatin (Figure 1.1-3). Removal of this protein results in a reduction in H3K4me3 levels arguing that its role may be to tether the COMPASS complex to specific regions of chromatin. Alternatively this loss in H3K4me3 may represent an essential role for *Spp1* in the enzymatic viability of this complex.

In the mammalian genome, H3K4 methylation seems to be a more complex affair. Whilst one SET complex appears to be responsible for the control of H3K4 methylation in yeast, multiple histone methyltransferases (HMTases) appear fill this role in the mammalian genome. Although it is unknown why such a number of H3K4 HMTases are required in higher eukaryotes, likely explanations are that the complexes are either functionally redundant or that each modify particular regions of the genome. Evidence exists to support the latter hypothesis as specific deletions and truncations of three different H3K4 HMTase genes in mice all give rise to distinct phenotypes (Yu, Hess et al. 1995; Glaser, Schaft et al. 2006). Mammalian H3K4 HMTases all contain a highly conserved SET methyltransferase domain, similar to those seen in the yeast *Set1* and *Drosophila* TRX proteins. The best characterised of these histone methyltransferases belonging to a family of proteins called the MLL proteins (Milne, Briggs et al. 2002; Wysocka, Myers et al. 2003; Yokoyama, Wang et al. 2004). This family contains the members MLL1, 2, 3 and 4 as well as Set1A and 1B (Figure 1.1-3) and exist in multi-protein complexes containing the mammalian homologues of the COMPASS complex found in *Saccharomyces cerevisiae* (reviewed by (Shilatifard 2008).

These multi-protein complexes share a common set of subunits to the *S.Cerevisiae* COMPASS complex such as the CPS homologues ASH2, WDR5 and RbPB5 (Wysocka, Myers et al. 2003; Lee and Skalnik 2005) . These proteins are essential for enzymatic activity, as removal of just one of the aforementioned subunits results in loss of activity *in vitro* and *in vivo* (Dou, Milne et al. 2006). Several of these subunits have been studied in greater detail such as the yeast CPS30 homologue, WDR5. This protein, as in yeast, contains WD40 repeats responsible for the formation of protein-protein interactions. Furthermore it has been shown to be responsible for the binding of H3K4 MTase complexes to histone tails already carrying the H3K4me3 modification (Wysocka, Swigut et al. 2005). This may result in the propagation of the H3K4me3 mark along chromatin.

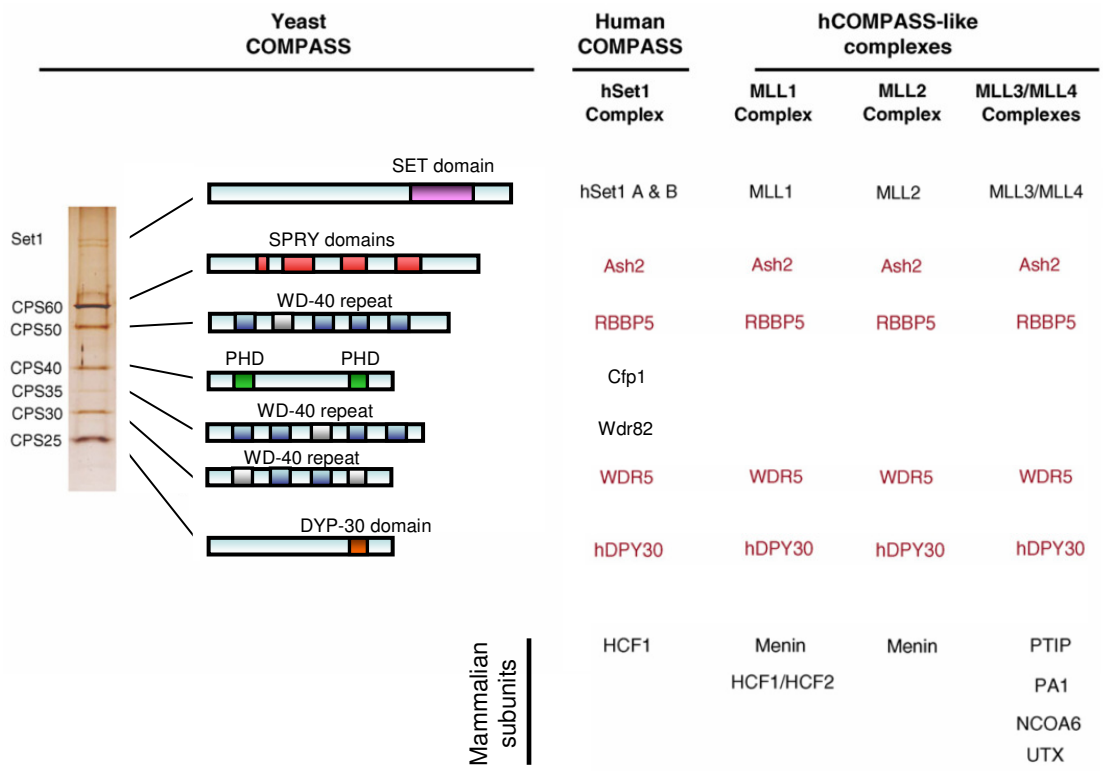

Figure 1.1-3. The Trithorax proteins in yeast and human. The components of the yeast COMPASS complex and the homologues in human. The core components are conserved between the human hSet1, Mll1, Mll2 &Mll3/4 complexes (purple text) with the catalytic SET domain protein as well as several other subunits varying between complexes (black text). Re-drawn from (Shilatifard 2008)

Although many of the subunits are shared between the mammalian H3K4 HMTase complexes, a few individual proteins vary. The Set1a and Set1b complexes are by and large identical in their protein content with the only difference being the catalytic methyltransferase present (either the Set1a or 1b protein). Presumably this allows for overlapping and non-redundant properties between complexes. Studies investigating the distribution of the Mll1 complex within cultured human cells found a tight correlation with sites of RNA polymerase occupancy (Guenther, Jenner et al. 2005). Mll1 was seen bound to 5,186 promoters which represent 23.3% of annotated protein coding promoters. Although the exact roles for the H3K4 HMTase complexes are not fully understood, each is thought to play an important role in the initiation and propagation of the H3K4me3 modification.

ii) Regulation of H3K27 methylation by the Polycomb group of proteins

The proteins responsible for the methylation of H3K27 residues were first discovered in *Drosophila* as proteins able to control the expression of genes responsible for the development of body segments (Lewis 1978). Subsequent work revealed that these proteins remodel the chromatin resulting in epigenetic maintenance of silenced gene state (reviewed by (Gil and Peters 2006; Sauvageau and Sauvageau 2008). In mammals these proteins, called the "polycomb proteins", form two discrete groups. The complex responsible for the initial methylation of H3 tails at K27, termed Polycomb repressor complex 2 (PRC2), is made up of three core components; Enhancer of Zeste (EzH2), Extra sex combs (Esc also known as Eed) and Suppressor of Zeste-12 (Su(z)12), along with several other proteins likely to have roles in directing the complex to specific loci (Figure 1.1-4).

The specific methylation of H3K27 residues occurs through the catalytic SET methyltransferase domain found in the EzH2 protein whilst chromatin binding is facilitated through interactions with the PHD zinc finger of Su(z)12 (Birve, Sengupta et al. 2001). The methylation of H3K27 tails leads to the attraction of a second complex of polycomb proteins, PRC1, which results in the compaction into heterochromatin (Figure 1.1-4).

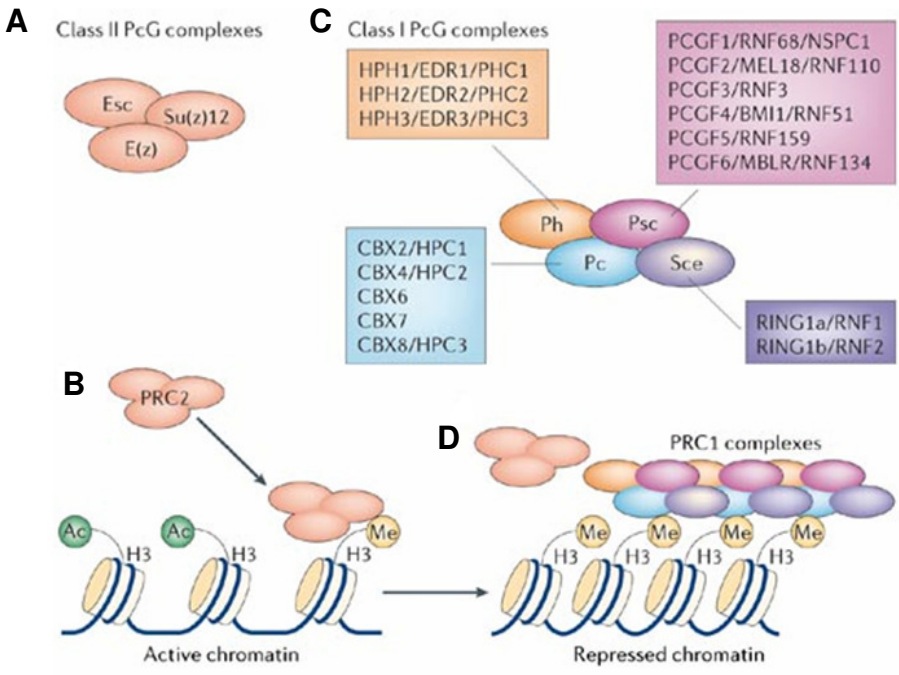

Figure 1.1-4. Mammalian Polycomb repressor complexes A. Class II polycomb complex (PRC2) is made up of three proteins, Esc, Su(z)12 and the catalytic subunit E(z) responsible for the methylation of H3K27 residues (**B**). **C.** PRC1 complexes are more complex in their formation than PRC2. Although consisting of four subunits, multiple forms of each exist as shown by the boxed lists. **D.** PRC1 complexes bind to the H3K27me3 modifications set by the PRC2 complex. These PRC1 complexes form compacted and repressed chromatin.
(Figure from (Gil and Peters 2006)

iii) Dynamic regulation of methyl groups from lysine 36 of histone H3

As H3K36me3 modified tails are involved in the progression of transcription through the gene body in conjunction with acetylation, this modification is somewhat dynamic (see 1.1.1 part iii above). It is therefore important that this modification can be removed as well as deposited. The protein Set2 interacts with RNAP II and specifically methylates lysine 36 on histone H3. Set2 is recruited to phosphorylated serine 2 residues on the RNAP II C-terminal domain facilitated by the Ctk1 protein (Xiao, Shibata et al. 2007). This in turn ensures that the processes of transcriptional elongation and H3K36 methylation are tightly coordinated. Once the transcriptional machinery has passed the H3K36 methylation must be removed. The removal of methyl groups is carried out by two classes of proteins; the LSD1 group and the Jumonji C domain proteins (JmjC). Individual members of these groups can

each demethylate specific lysine residues such as the affinity of Kdm5s for H3K4me3 and Kdm3a for H3K9me3 demethylation (Klose, Kallin et al. 2006; Ingvarsdottir, Edwards et al. 2007). Recent work has revealed the Kdm2a protein as the demethylase responsible for the removal of methylation from H3K36 residues (Blackledge, Zhou et al. 2010)

iv) Regulation of acetylation across the histone H3 tail

The acetylation and deacetylation of arginine and lysine residues on histone tails is thought to be a dynamic process tightly linked to transcription (Hebbes, Clayton et al. 1994). This dynamic nature requires the coordinated addition and removal of acetyl groups from histone tails which is a process carried out by a group of histone acetyl transferases (HATs) and histone deacetylases (HDACs) (Wang, Zang et al. 2009). Three major families of HATs have been found in mammals: the Gcn5 related acetyltransferases (GNATs) which specifically target H3 tails, the MYST family of HATs which target the H4 tails and the CBP/p300 family which can target both tails. (Glozak, Sengupta et al. 2005). Although functionally similar, structural analysis reveals that each group of HATs are distinct from one another (Vetting, LP et al. 2005). The HDAC proteins fall into four main groups based on their catalytic activity. Type I, II and IV HDACs have a related mechanism for deacytlation whilst the third group, type III, require the cofactor nicotinamide adenine dinucleotide (NAD) (Kurdistani and Grunstein 2003; Vaquero, Sternglanz et al. 2007). Thus it can be seen that the regulation of acetylated histone tails is a highly complex process facilitated by multiple protein complexes and one which must be coordinated tightly in order to facilitate the transcription of specific genes.

1.2 Epigenetic modification of the DNA

The epigenetic mechanisms discussed so far have all involved modifications to the tails of the abundant DNA binding proteins; the histones. However the DNA can also be directly modified by the incorporation of a methyl or hydroxymethyl group (Figure 1.2-1). To date the majority of work has focused around the methylation of DNA and although the distribution of this covalent modification is seen to vary between many organisms the functional implications remain the same.

1.2.1 DNA methylation across organisms

i) DNA methylation in bacteria

In *E.coli*, methylation can occur at the 6th nitrogen of the adenine base and 4th nitrogen and 5th carbon of the cytosine base (Noyer-Weidner and Trautner 1993). Interestingly this modification appears to act as a defence system against invading bacteriophages as exogenous sequences lacking such methyl marks are vulnerable to degradation by methylation sensitive restriction enzymes encoded by the bacteria (Meselson, Yuan et al. 1972; Bestor 1990). Further to this role, methylation has also been implicated in the regulation of transcription as removal of DNA methylation in bacteria led to the overexpression of many genes (Low, Weyand et al. 2001).

ii) DNA methylation in fungi

The presence and distribution of methylated DNA varies widely between members of the fungi family. This epigenetic modification has not been detected in the ascomycetes *S.cervisiae and S.pombe* (Proffitt, Davie et al. 1984) however high levels of methylated cytosine (methylated at the 5th carbon) are found in the filamentous fungi *Neurospora crassa* at single cytosine bases (Selker, Tountas et al. 2003). Furthermore these 5 methyl cytosine (5meC) nucleotides are seen to associate specifically within transposable elements and removal of this methylation results in increased transposition.

iii) DNA methylation in plants

Methylation of the DNA is also found in plants, where as much as 6% of the *Arabidopsis* genome contains the 5meC nucleotide (Bender 2004). In contrast to the simpler organisms described above this modified base can exist within the context of CpG, CpNpG, and CpNpN sequences where N represents any nucleotide but guanine (Finnegan, Genger et al. 1998). The distribution of this modification throughout the *Arabidopsis* genome reveals that the majority of gene bodies are devoid in 5meC, whilst a minority of promoters contain this mark and tend to exhibit tissue specific expression profiles. Once again, removal of DNA methylation results in the expression of transposons

(Zhang, Yazaki et al. 2006) suggesting that the major role for DNA methylation in plants is similar to the genome defence roles employed by bacteria and fungi.

iv) DNA methylation in invertebrates

DNA methylation in invertebrates such as *Ciona intestinalis* is a more complex affair. Generally methylation patterns are mosaic; occurring in domains interspersed with non-methylated domains. These non-methylated regions tend to co-localise to promoters, intergenic DNA sequences as well as transposons (Suzuki, Kerr et al. 2007). As such, methylation appears to be targeted to the bodies of genes and it is thought that the role of DNA methylation here is to prevent the spurious initiation of transcription from within a gene. Once again, DNA methylation is intrinsically linked to inaccessible and silenced chromatin. It must be noted however that several invertebrate genomes are found to be completely devoid of DNA methylation such as the nemadote *C.elegans* or contain extremely low levels, as is the case with the fruit fly *Drosophila* (Bird 2002). As these organisms still contain defined regions of histone tail modifications and are able to silence transcription efficiently this argues that an alternative mechanism for gene silencing separate to DNA methylation exists in these organisms. Quite why DNA methylation is absent from *C.elegans* and *Drosophila* is not known.

1.2.2 DNA methylation in mammals

i) CpG as a signalling molecule

In the mammalian genome, epigenetic modification of the DNA occurs at cytosine bases within the dinucleotide CpG (Figure 1.2-1). The most abundant form of this dinucleotide (representing around 70% of methylated bases) is modified through the incorporation of a methyl group onto the fifth carbon in the cytosine base (5 methyl cytosine). As in other organisms, this methylated cytosine base tends to be found within compacted chromatin and has links to the silencing of associated genes and transposable elements. Whether or not the methylation of DNA is directly responsible for these events or is a secondary event of heterochromatic compaction is as yet unknown.

A second state in which the CpG dinucleotide is found contains a modified 5-hydroxy methyl cytosine (5hmC) base. Similar to the methylation of cytosine bases, a hydroxymethyl group is present on the

5th carbon on the cytosine base. Although first characterised in the bacteriophage in 1952 (Wyatt and Cohen 1952) it was only recently found in mouse and human (Kriaucionis and Heintz 2009) and as yet no known hydroxymethyl CpG (hmCpG) binding proteins have been identified. To date 5hmC has only been found in a few cell types such as the purkinje neurons in the brain and embryonic stem cells. Whether or not this modified base represents a unique signalling platform or is simply an intermediate in demethylation processes is as yet unknown.

Although the majority of CpG dinuleotides exist in a methylated form, around 30% is non-modified and is of particular interest due to its unique distribution throughout the genome. These non-methylated cytosine bases tend to be found over 1kb stretches of CpG rich DNA, often at discrete loci such at the 5' ends of genes (Bird, Taggart et al. 1985; Bird 1986). In contrast to methylated cytosine, this non-modified form of the CpG dinucleotide is associated with the promoters of active genes and regions of the genome in which the chromatin is decondensed.

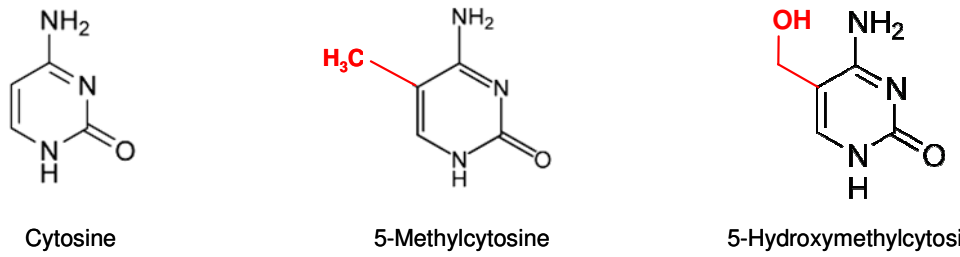

Figure 1.2-1. Cytosine bases within CpG dinucleotides exist in three forms. The cytosine base can exist in three distinct states within the dinucleotide CpG in the mammalian genome. Around 30% of cytosine is found in a non modified state which tends to occur in regions of high CpG density. The majority of CpG (~70%) exists in a modified form in which a methyl group is found at the carbon 5 position of the DNA base. In contrast to these forms of cytosine, relatively little is known regarding cytosine containing a hydroxymethyl group however this base may represent an intermediate between methylated and non-modified cytosine.

ii) Distribution of DNA Methylation

Although the majority of CpG dinucleotides exist in a methylated state, they do so over a large fraction of the genome (~98%) at low densities around 1 per 100 base pairs of DNA (Bird 1980).This deficiency in the density of methylated CpG dinucleotides is largely due to spontaneous deamination

of the methylated cytosine base. As methyl CpG is relatively unstable, spontaneous deamination of this modified dinucleotide through a hydrolysis reaction converts the methyl cytosine into thymine. As the resulting conversion is not efficiently repaired by the DNA repair machinery this results in an under-representation of this dinucleotide. Although deficient in CpG density, regions of high meCpG are typically found associated with transposable elements, intragenic regions and repeat sequences (Weber, Davies et al. 2005; Eckhardt, Lewin et al. 2006). In contrast to these methylated CpGs, the remainder of the genome (<2%) is rich in non modified CpG dinucleotides (typically 1 CpG per 10 base pairs) at discrete loci. These unique regions, termed CpG islands or CGIs will be explained in greater depth below (see chapter 1.3)

Methylation of the cytosine base has been associated with gene silencing events and the formation of localised heterochromatin across a host of organisms. Early experiments revealed that the gene encoding adenine phosphoribosyltransferase (*aprt*) was silenced by methylation of CpG dinucleotides over its promoter region upon transfection into cultured mammalian cells (Stein, Razin et al. 1982). Additional experiments revealed that removal of DNA methylation through the use of the drug 5-azacytidine (5 Aza-C) resulted in the restored expression of genes on the inactive X chromosome which were previously identified as silenced by the methyl mark (Mohandas, Sparkes et al. 1981; Wolf, Jolly et al. 1984). Precisely how the addition of a methyl group results in the silencing of underlying genes is still subject to some debate. One possible explanation is that this silencing may arise through the direct steric hindrance of transcription factors and binding proteins by the methyl group itself. Alternatively meCpG can act as a signalling module for the binding of a series of proteins which contain methyl binding domains (MBDs). Many of these proteins can either directly, or indirectly influence local chromatin modifications resulting in a transcriptionally repressive environment.

iii) DNA methylation as a heritable epigenetic mark

The methylation of cytosine nucleotides is a heritable epigenetic mark in that it is copied over onto newly synthesised strands of DNA upon replication. Early work on establishing the methylation patterns in mammals was carried out through the use of methylation sensitive restriction enzymes

where the methyl mark protects the CpG site from digestion. In the context of double stranded DNA it was found that the cytosine within a CpG dinucleotide would be either be methylated on both strands or unmethylated on both, fitting with a model of a transmissible methylation mark (Bird 1978). The fact that the patterns are faithfully copied over during cell division give rise to a mechanism of cell memory which is essential for viability (Li, Bestor et al. 1992).

1.2.3 Setting up the DNA methylation pattern: the *de novo* methyltransferases

DNA methylation patterns in mammals are set up during gametogenesis and again in embryogenesis (Reik, Dean et al. 2001; Reik 2007; Weber and Schubeler 2007). Initially the genomes of mature sperm and egg cells contain high levels of methylated DNA, equivalent to those found in somatic cells (Bestor 2000). During the early stages of development the levels of DNA methylation drop precipitously. Analysis of global DNA methylation levels in the fertilised egg using an antibody against 5meC revealed that one chromosome appeared remarkably different from the other in its methylation content (Rougier, Bourc'his et al. 1998). Closer inspection of the zygote prior to the fusion of the maternal and paternal pronuclei revealed that although both were initially highly methylated, there was a dramatic loss of methylation from the paternal genome a few hours after fertilisation (Mayer, Niveleau et al. 2000; Oswald, Engemann et al. 2000). Intriguingly, this loss of methylation occurs in the absence of DNA replication and as such appears to be an example of an active demethylation, possibly facilitated through an as yet unidentified DNA demethylase enzyme. In contrast, the maternal genome also loses DNA methylation during early embryogenesis. However this loss is a passive event, occurring as a result of cell division in the absence of a DNA methyltransferase enzyme. In total more than half of the initial DNA methylation levels are lost at this stage prior to re-methylation once implantation has occurred.

The enzymes responsible for the methylation of DNA can be separated into two groups; those which can set up the patterns of methylation during development and those which maintain it in differentiated cells. So far four mammalian DNA methyltransferases (Dnmts) and one specific cofactor have been discovered all of which share at least a large part of a conserved catalytic domain

responsible for the methylation of the DNA (Figure 1.2-2). Two of these Dnmts have been shown to be essential for the setting up of the methylation patterns after the early developmental global demethylation events and as such are termed the "*de novo* Dnmts". This group consists of the enzymes Dnmt3a, 3b along with the 3a cofactor Dnmt3L. These *de novo* methyltransferases were first discovered through the observation that non-methylated DNA would remain in this non methylated state after multiple cell divisions. However non-methylated transgenes introduced into mouse preimplantation embryos were found to become stably methylated in the adult mouse (Jahner, Stuhlmann et al. 1982). Disruption of the gene for the previously discovered DNMT1 enzyme did not interfere with this de novo methylation and as such was excluded as a potential *de novo* methyltransferase. A search of EST databases using the catalytic domain of DNMT1 brought up three potential candidates, DNMT2, 3a and 3b (Okano, Xie et al. 1998). Further work on DNMT3a and 3b revealed that neither of these proteins were able to methylate hemimethylated DNA (unlike DNMT1), implying that these proteins act on non-methylated templates (Okano, Xie et al. 1998). Finally, disruption of both of these genes in embryos and ES cells led to a genome wide loss of methylation, leading to the conclusion that these complexes were indeed the *de novo* methyltransferases (Xie, Wang et al. 1999).

Subsequent work has revealed that the protein DNMT3L is an important cofactor for DNMT3a and is required for the specific methylation at imprinted loci (Hata, Okano et al. 2002). Dnmt3L closely resembles the carboxy terminus of the other *de novo* methyltransferases but lacks critical sequence motifs required for catalytic activity (Bourc'his, Xu et al. 2001). Recent studies have identified the crystal structure of this interaction between Dnmt3a and the cofactor 3L revealing that the periodicity of methylation from this complex was in the range of 8 to 10 base pairs (Ooi, Qiu et al. 2007) which is in agreement with earlier biochemical predictions. Furthermore this study revealed that DNMT3L cannot bind histone H3 tails in which the 4^{th} lysine is methylated. As H3K4me3 is known to be linked to actively transcribed genes this may link the methylation of the DNA to histone modifications resulting in gene silencing. Additionally DNMT3L is able to bind to histone H3 tails modified by methylation at K27 and K9, both marks of gene silencing (Ooi, Qiu et al. 2007). These specific

interactions with histone tails might go some way to explain how such Dnmt complexes are targeted to specific loci thus setting up methylation profiles during development.

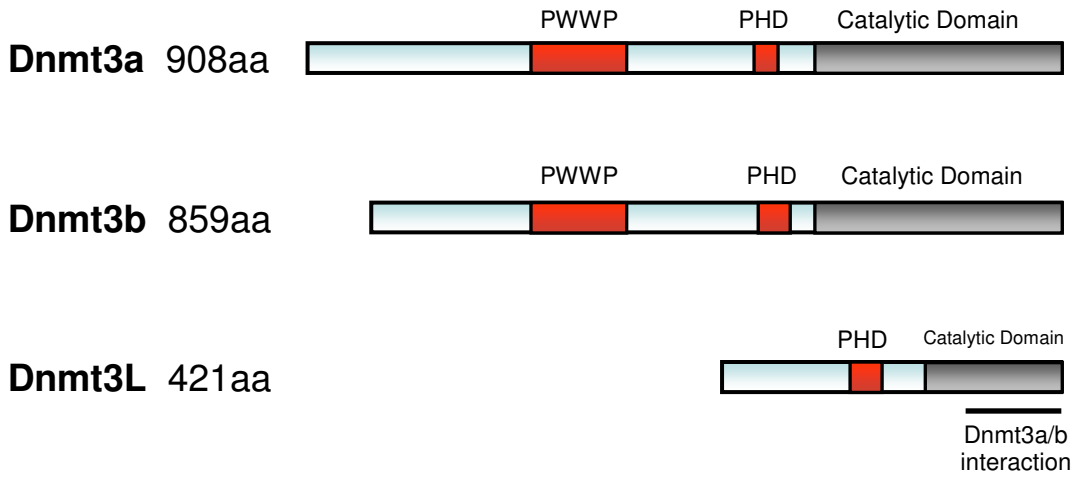

Figure 1.2-2. The *de novo* methyltransferases. The two *de novo* Dnmts aligned with the Dnmt3L cofactor. Both Dntm3a and b contain PWWP domains which aids binding to DNA. All three proteins contain a chromatin binding PHD domain along with catalytic domains (grey box) although the catalytic domain of Dnmt3L lacks certain motifs leading to loss of catalytic ability. Shown is the region with which Dnmt3L interacts with Dntm3a and b.

1.2.4 Maintaining the DNA methylation pattern

As DNA methylation has roles in the regulation of transcription it is essential that the methylation patterns are faithfully propagated with every cell division. The copying of this pattern is carried out by another group of DNA methyltransferases, the maintenance methyltransferases. In fact, the first DNA methyltransferase discovered was the maintenance methyltransferase Dnmt1 (Bestor and Ingram 1983). This methyltransferase progressively copies the methylation pattern from the parental strand onto newly synthesised DNA strands allowing faithful propagation during replication. Dnmt1 preferentially binds to hemimethylated DNA (Yoder, Soman et al. 1997) such as that found at the replication forks. This preference for hemimethylated DNA is brought about through interactions between the amino terminal domain of this protein and the DNA. Removal of the N terminal region of Dnmt1 abolishes the specificity for hemimethylated DNA whilst increasing the normally limited *de novo* activity of the protein (Bestor 1992; Chuang, Ian et al. 1997). Not only does this affinity for

hemimethylated DNA help target Dnmt1 to sites of ongoing replication, direct interaction with the proliferating nuclear antigen (PCNA) protein ensures that Dnmt1 tracks along with the replication fork (Chuang, Ian et al. 1997) resulting in the copying of methylation patterns as replication is taking place.

Interestingly Dnmt1 contains a CXXC domain within the N terminal half of the protein. Proteins containing this domain have been found to bind specifically to non-methylated CpG rich sequences (Voo, Carlone et al. 2000). This domain, absent in the *de novo* dnmts, may have a role in the hemimethylated binding exhibited by Dnmt1 as hemimethylated sites will indeed contain large amounts of non methylated CpGs.

Dnmt1 is active only within the nucleus with targeting of this protein accomplished by a nuclear localisation sequence (NLS) (Figure 1.2-3) within the N terminus. Interestingly, modulation of this NLS can give rise to a further level of Dnmt1 control. Nuclear targeting does not occur during the early stages of embryonic development resulting in accumulation of Dnmt1 outside of the nucleus. This nuclear exclusion overlaps with passive global demethylation events in the female pronucleus of preimplantation embryos (Oswald, Engemann et al. 2000).

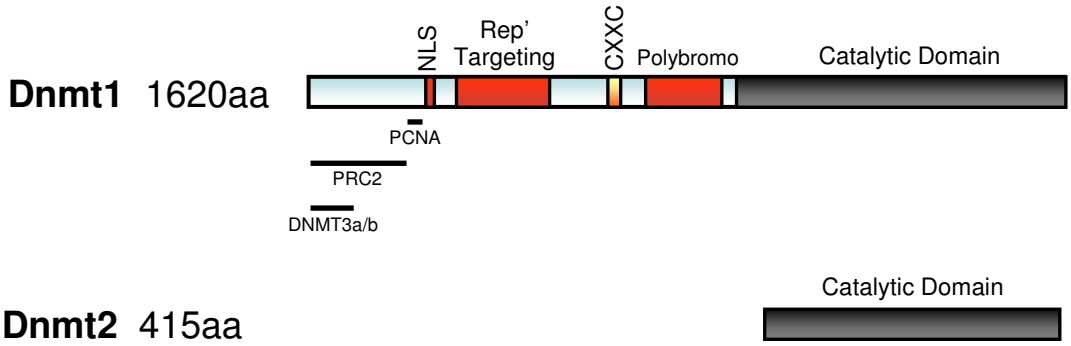

Figure 1.2-3. The maintenance DNA methyltransferases. Dnmt1 and Dnmt2 domain architecture and interactions. The highly conserved catalytic domain shared by both proteins is represented by the grey box. In addition to the non-methyl binding CXXC domain (yellow box), Dnmt1 contains a nuclear localisation signal (NLS, red box), a region important in replication targeting (rep' targeting, red box) and a polybromo domain involved in protein-protein interactions. In addition to these domains interactions with PCNA, PRC2 and Dnmt3a/b are represented underneath by black bars.

The second DNA methyltransferase discovered through the conserved catalytic domain was Dnmt2 (Figure 1.2-3) although this protein appears to lack methyltransferase activity (Schaefer and Lyko 2010). Although the structure of Dnmt2 is closely related to that of Dnmt1 (Dong, Yoder et al. 2001), recent studies have found that a tyrosine residue found only in Dnmt2 and not Dnmt1 may prevent it from binding to DNA efficiently (Goll and Bestor 2005). However rather than act as a DNA methyltransferase this protein was found to instead methylate RNA, specifically to methylate position 38 in aspartic acid transfer RNA (mC38 $tRNA^{Asp}$). Following this finding, Dnmt2 is now more commonly referred to as tRNA aspartic acid methyltransferase 1 (TRDMT1).

1.2.5 Mediators of methylation: The MBD proteins

The dinucleotide meCpG can lead to long term stable gene repression; however this modification on the DNA must be translated into an overlying epigenetic change. As explained above this is likely to be brought about be either steric hindrance or through the recruitment of methyl binding proteins along with associated transcriptional repressors. Indeed the latter hypothesis has been the more thoroughly investigated. A group of methyl specific binding proteins have been identified which contain a highly conserved methyl binding domain or MBD (Figure 1.2-4). In addition, many of these proteins have been shown to interact with transcriptional repressor complexes (Bird and Wolffe 1999). These methyl binding proteins were first identified through band shift assays using methylated probes as bait. Using this method the protein complex MeCP1 was discovered (Meehan, Lewis et al. 1989). The first purified protein to be identified however was Methyl-CpG binding protein (MeCP2). Database searches for proteins containing similar methyl binding domains to that of MeCP2 gave rise to a further four proteins which together make up the MBD family of proteins (Figure 1.2.4) (Hendrich and Bird 1998; Bird and Wolffe 1999).

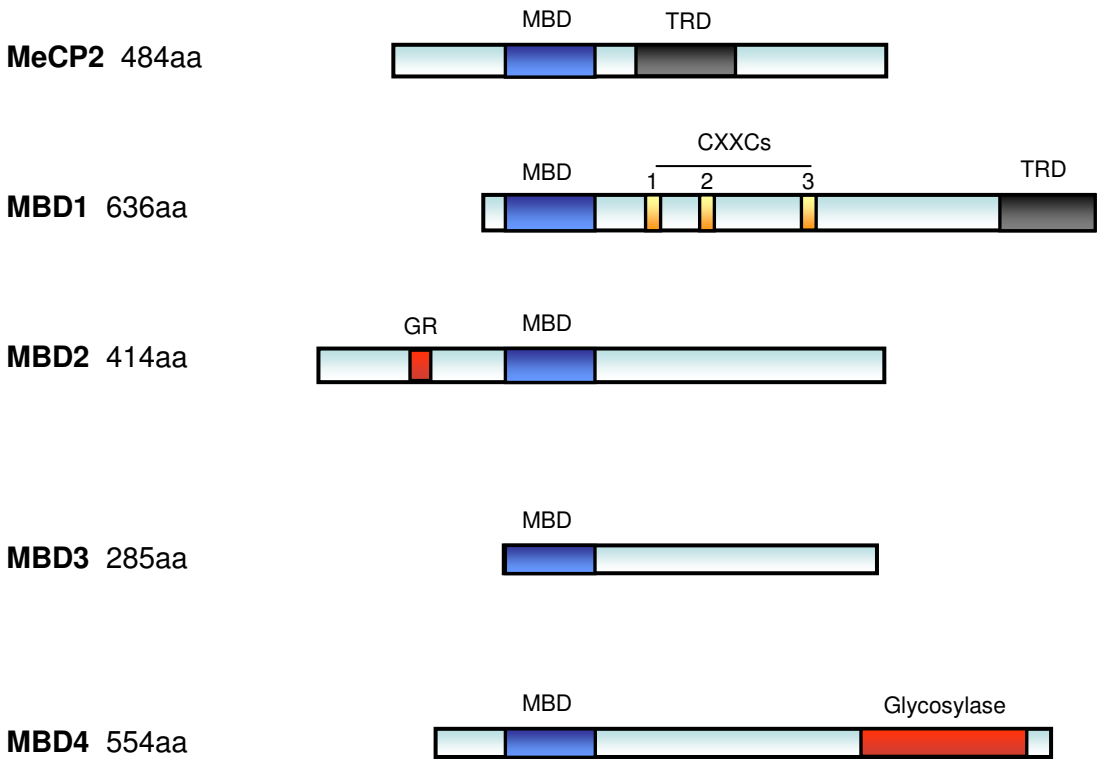

Figure 1.2-4. The MBD family of proteins. The MBD group of proteins are aligned by their methyl binding domain (blue). Two members, MeCP2 and Mbd1 contain a transcription repression domain (TRD, grey boxes). The protein Mbd1 also contains 3 non-methyl CpG binding domains (CXXC, yellow boxes). MbD4 is involved in DNA repair and as such contains a glycosylase domain (red box)

The first of the MBD proteins to be discovered, MeCP2, was shown to contain both an MBD domain for specific meCpG binding as well as a carboxy terminal transcription repression domain (TRD). This domain interacts with the Sin3a complex which contains histone deacytalase activity (Nan, Ng et al. 1998). As such, recruitment to methylated CpGs results in removal of the acetyl groups from histone tails and as these modifications are essential for formation of a decondensed chromatin state, transcription is subsequently silenced. The levels of this MeCP2 protein are highest within the neurons of the brain and this is consistent with the findings that MeCP2 deficiency results in cases of severe mental retardation known as Retts syndrome (Nan and Bird 2001). Recent work in the Bird laboratory focussing on the MeCP2 rich neuronal cells of the brain has shown that MeCP2 is bound globally to methylated DNA and not at specific methylated loci as once thought (Skene, Illingworth et al. 2010). Further to this, removal of this protein resulted in globally elevated levels of histone

acetylation linking modification of the DNA to changes in the chromatin state through this methyl binding protein.

Another protein containing an MBD domain, Mbd1, interacts with the histone H3 lysine 9 methyltransferase (H3K9 HMTase) SetDB1 (Hendrich and Bird 1998). As the H3K9me3 modification is known to be associated with gene silencing events it is another example of the methyl binding proteins recruiting transcriptional repressors to methylated CpG rich sequences. This recruitment is thought to occur in parallel with replication due to the transient interaction between MBD1 and the chromatin assembly factor (CAF) during S phase (Reese, Bachman et al. 2003). MBD1 is of particular interest as not only does it contain a domain for specific methyl CpG binding, it also contains three CXXC domains with particular affinities for non methylated CpG dinucleotides. Investigations into the *in vitro* and *in vivo* binding properties of these three CXXC domains reveals that only the most 3' of these is able to bind to non methylated DNA (Jorgensen, Ben-Porath et al. 2004). Closer inspection of the sequences of these CXXC domains reveals that CXXC-3 contains high levels of conservation with previously verified functional CXXC domains, whereas the other two domains exhibit functionally important changes rendering these domains unable to bind non methylated CpGs. However the role of the functional CXXC domain of MBD1 is still largely unknown. One possible explanation is that during replication, MBD1 can bind to both the methylated and newly synthesised (and thus non-methylated) strand at the replication fork. This assumes that the copying of methyl groups onto the daughter strand by Dnmt1 occurs after the silencing histone modifications such as H3K9me3 have been deposited (Sarraf and Stancheva 2004).

Another protein containing an MBD is the protein Mbd2. The MBD2 protein has been purified as part of the nucleosomal remodelling complex (NuRD) complex which also binds to HDACs 1 & 2, thus linking DNA methylation to changes in histone modifications (Zhang, Ng et al. 1999) Studies have shown that fibroblast cell lines derived from *Mbd2* null mice exhibit a reduced ability to silence a methylated reporter gene highlighting a potential role in gene silencing (Hendrich, Guy et al. 2001). Further roles for Mbd2 have arisen through the finding that *Mbd2* null mice have striking defects in the production of cytokines (Hutchins, Mullen et al. 2002). This appears to be due to the fact that

Mbd2 binds to the promoter of the *Il4* gene responsible for the formation of TH2 T-helper cells involved in the immune response. Upon stimulation by antigen presenting cells, the naïve T helper cells express *Il4* during differentiation into TH2 cells which occurs through the loss of Mbd2 binding from the *Il4* promoter. As such Mbd2 appears to be a regulator of transcription however to date very few promoter specific binding sites have been found for this protein.

These MBD proteins represent a mechanism of cross talk between methylated DNA sequences and the local chromatin structure. As the methylated CpG dinucleotide is acting as a signalling module for the binding of a specific set of proteins it will be interesting to determine whether or not similar mechanisms of binding are occurring at the non-modified CpG dinucleotides.

1.3 CpG islands

Although the majority of CpG dinucleotides in the mammalian genome exist in a methylated form, around 30 % is non-modified and plays a crucial role in the control of gene expression and localised chromatin state. Although the majority of non-methylated CpG dinucleotides are found dispersed across the majority of the genome a proportion (~8%) are found at discrete loci with a frequency of around one CpG per 10bp (Figure 1.3-1) (Bird, Taggart et al. 1985). Not only are such sites unique in the regards that they exist within a largely methylated genome but they tend to co-localise with the 5' ends of genes (Figure 1.3-1). These CpG rich regions were first discovered as fractions of the genome which were cleaved unusually frequently by methyl sensitive restriction enzymes such as HpaII (Cooper, Taggart et al. 1983). This enzyme only cuts the sequence CCGG when unmethylated resulting in the digestion of the non-methyl CpG rich loci into many small fragments which were subsequently termed "HpaII tiny fragments" (HTFs). Cloning of these HTFs revealed that these fragments were derived from the 1kb stretches of non-methyl CpG rich DNA we now know today as the CpG Island or "CGI" (Bird, Taggart et al. 1985).

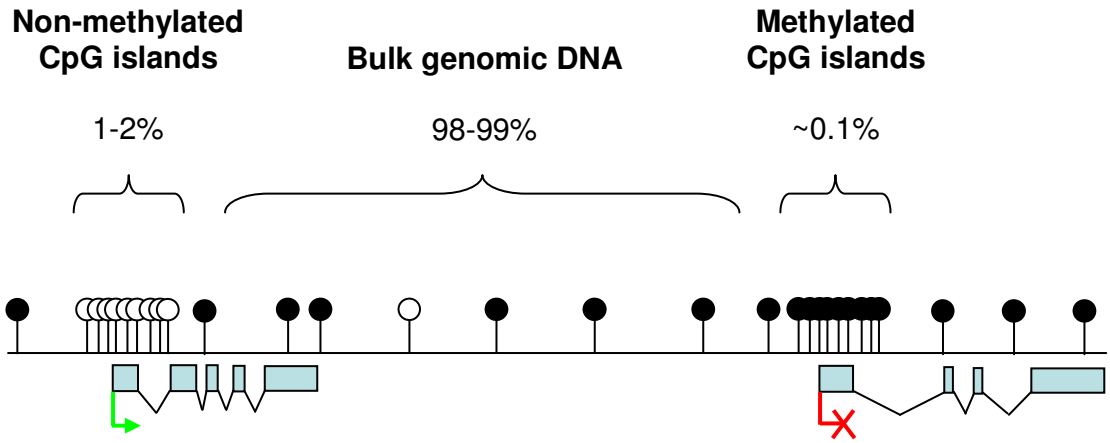

Figure 1.3-1. Distribution of CpG dinucleotides within the mammalian genome. The distribution of CpG states are shown for a typical mammalian cell. Percentage values represent proportion of total genomic DNA. White lollypops represent non methylated CpGs whilst black lollypops are methylated. Exons are represented underneath as blue boxes. Transcriptional states of these genes are represented by the green arrow (active) or the red cross (inactive). The majority of the genome (98-99%) contains low densities of CpG dinucleotides, mainly methylated. 1-2% however correspond to regions of high CpG density. These regions are generally non-methylated and associated with the 5' ends of genes (CpG islands). A minority (~0.1% of total CpG dinucleotides) of these islands are methylated, corresponding with a loss of transcription. This figure does not represent the 40% of gene promoters which lack a CGI, most of which are tissue specific in function.

1.3.1 Characterisation of the CpG Island

Mapping studies across the promoter region of the gene for alpha 2 (type 1) collagen in chicken cells provided an early insight in the unique properties of CpG island DNA. Although the CpG rich regions over promoters had not been identified as the entities now known as CGIs it was noted that the CpG rich sequence remained unmethylated regardless of the transcriptional state of the gene whilst the gene body remained methylated. (McKeon, Ohkubo et al. 1982) The majority of the early work on CGIs revolved around the use of digestion by methyl sensitive restriction enzymes. The human CGI fraction was first liberated in this way as digestion with HpaII or HinP1 was sufficient to separate the high molecular weight bulk DNA from the highly digested and thus small fragments of non-methylated CGI DNA. When carried out within the context of chromatin the resulting digestion profiles varied slightly, indicating that the presence of nucleosomes occludes many of the potential methyl and non methyl CpG cut sites (Antequera, Macleod et al. 1989; Tazi and Bird 1990). Purification of these CGI fragments along with their associated histones revealed high levels of H3 and H4 acetylation with depleted levels of the linker histone H1 (Tazi and Bird 1990). These early investigations outlined

CGIs as not only unique in their DNA sequence composition and cytosine methylation levels but in their chromatin modifications and levels of chromatin compaction.

CGIs are found to correspond to the promoters of around 60% of human genes (Lander, Linton et al. 2001; Saxonov, Berg et al. 2006; Weber and Schubeler 2007). Although the majority of these are highly active housekeeping genes, many genes with restricted tissue expression patterns also contain a CGI over their 5' end. Tissue specific genes lacking CGIs tend to contain TATA boxes for the attraction of transcriptional machinery (Yang, Bolotin et al. 2007). These TATA boxes rely heavily on the coordinated presence of transcriptional activators and as such represent a more highly regulated mode of transcriptional control.

1.3.2 Defining and Mapping CpG islands

As the methods for CGI detection have developed, so too has our understanding of the distribution of islands. Initial predictions were made by extrapolating the results of observed CGI frequencies across the entire genome resulting in a predicted 45,000 Islands (Antequera and Bird 1993). This prediction was in fact very close to the number of Islands predicted through bioinformatic analysis of genomic sequences using a set of predefined parameters. Typically the criteria used for the identification of CGIs according to standards defined by Takai and Jones is to search the genome for DNA sequences with a base composition of Cytosine and Guanidine ("C+G") greater than 50%, a CpG observed over expected ratio [o/e] of at least 0.6 (Takai and Jones 2002) over a length of either a 200 bases ("NCBI - relaxed") or 500 bases ("NCBI-strict"). As such depending on the parameters used, relaxed or strict, the number of CGIs identified changes with the stringency of the analysis. Analysis using the "NCBI-relaxed" criteria gives over 300,000 predicted islands, however nearly 90% of these islands were unable to be identified through subsequent CGI mapping studies (Illingworth, Kerr et al. 2008). By contrast the "NCBI-strict" data set predicts around 24,000 islands. Although this generates a far more convincing data set, subsequent CGI mapping studies failed to identify a quarter of these predicted islands. These findings highlight the problems with the computer based algorithm approach to CGI determination.

As computational prediction methods were not sufficient to generate complete CGI data sets more accurate studies were devised utilising the affinity of a protein domain for non-methylated CpG dinucleotides. This "CXXC domain", so called due to conserved cysteine residues forming a DNA binding zinc finger, has been shown to harbour an affinity for non methylated CpG rich probes *in vitro* (see chapter 1.4)(Voo, Carlone et al. 2000). As such, columns containing the CXXC-3 domain from the Mbd1 protein were used to select for non-methylated CpG rich fragments of DNA such as those found at CGIs in a process termed "CAP" (CXXC affinity purification). The resulting purified DNA fragments were then hybridised to arrays to map the locations of such CGIs. Pools from human whole blood DNA gave rise to 25,200 unique CGIs (Illingworth, Kerr et al. 2008), 76% of which were within 1.5kb of the TSS of an annotated gene. Interestingly only half of these CGIs were seen directly over the transcriptional start sites of genes. Previous to this finding CpG islands were thought to be primarily promoter specific however this finding along with more recent studies using high resolution whole genome sequencing technology (Illingworth, Gruenewald-Schneider et al. 2010) reveals a more complex distribution of CGIs within the mammalian genome which be broken down into three classes; annotated promoter CGIs, intra-genic CGIs and inter-genic CGIs. The roles of such non-promoter CGIs are as yet unknown but as these loci are often found associated with RNAPII one cannot rule out the possibility that they represent either non-annotated protein coding genes or non-coding RNA (ncRNA) genes.

1.3.3 Origins and maintenance

Non-methylated CGIs are exceptions to the global CpG methylation patterns seen throughout the majority of the genome. This has lead to these discrete loci becoming the focus of many studies aiming to determine how CGIs are both set up during development and how they are subsequently maintained in this non-methylated state in differentiated tissues.

CGIs as unmethylatable entities

One of the simplest ways to explain how CGIs remain non-methylated in a highly methylated genome is that these sequences have some property which makes them refractory to incorporation of methyl

cytosine. This hypothesis however is unlikely when one considers the fact that a fraction of islands can be methylated and silenced between differentiated cell types as well as being widely methylated on the inactive X chromosome (Park and Chapman 1994). Methylation of islands also occurs in many cancers leading to the silencing of certain genes (reviewed by (Bird 1996), therefore the underlying DNA sequences at CGIs are indeed capable of becoming methylated.

CGIs as regions of abnormal DNA metabolism

An alternative view is that non-methylated CGIs represent regions of the genome which are affected by abnormally high levels of DNA metabolism such as sites of origins of replication or regions of high DNA repair. These loci would result in high levels of turnover of the underlying DNA bases during replication firing or DNA repair which could result in the removal of methylated cytosine bases. However, thus far only limited and predominantly speculative studies have been carried out implying a link between origins of replication and CpG islands sites (Antequera and Bird 1999). Furthermore, how these putative sites of abnormal DNA metabolism would result in the exclusion of the Dnmt proteins is also unclear.

Active demethylation

An attractive model and one which has gained much support recently is that CpG islands are actively demethylated resulting in the removal of methyl groups from specific sites by an as yet unidentified enzyme. As a set of enzymes responsible for the methylation of DNA (Dnmts) have been identified it is widely believed that enzymes might exist that facilitate the opposite reaction. Indeed this is the case with many enzymes including the histone methyltransferases and demethylases. Studies in *Arabidopsis* have identified the candidate demethylase enzyme *Ros1* (Agius, Kapoor et al. 2006). This protein contains a glycosylase domain which has roles in base excision during DNA repair. During base excision repair, the damaged base is flipped out from the DNA double helix before cleavage of the N-glycosidic bond (Liu, Prasad et al. 2007). It is predicted that this mechanism may be responsible for the active removal of 5meC bases from targeted sites. Interestingly loss of *Ros1* leads to global hypermethylation with associated gene silencing events whilst overexpression in transgenic plants

results in reduction of 5meC levels along with increased expression of target genes (Agius, Kapoor et al. 2006). These findings demonstrate the likely existence of active demethylation at least in plants.

Although a multitude of candidate proteins have been proposed, no mammalian demethylase has yet been convincingly identified. For some time the methyl binding protein MBD4 was thought to be a potential DNA demethylase due to the presence of a glycosylase domain towards the C terminus of the protein. This protein together with thymine DNA glycosylase (TDG), is responsible for the removal of thymine bases from T:G mismatches produced through deamination of meCpG dinucleotides (Yoon, Iwai et al. 2003). However removal of this protein does not affect global levels of methylation but instead leads to a three to four fold increase in mutations at CpG sites implying that MBD4s main role is that of a DNA repair protein and not that of a demethylase. The protein GADD45a was also highlighted as a potential DNA demethylase (Barreto, Schafer et al. 2007) however subsequent studies reveal that this protein was unable to demethylate methylated reporter plasmids even when overexpressed (Jin, Guo et al. 2008).

As the search for a demethylase proved more difficult than first anticipated, attention turned to the methyltransferase enzymes themselves for further clues. Recent studies by two groups argue that the DNA methyltransferases responsible for the setting up of the methylation profiles, Dnmt3a and 3b, are also those responsible for the demethylation at CpG islands (Kangaspeska, Stride et al. 2008; Metivier, Gallais et al. 2008). These studies reveal that cyclical methylation of CpG dinucleotides occurs at the active promoter of the estradiol-estrogen receptor (E_2-ERα) and that this coincides with the recruitment of the *de novo* methyltransferases to this site. These Dnmts are thought to be involved in the deamination of meCpG to TpG resulting in loss of methylation. However the efficiency of the Dnmt deamination reaction is likely to be insufficient to facilitate the observed levels (Ooi and Bestor 2008). The most recent protein to be linked with roles as a DNA demethylase is that of the ten-eleven translocation-1 (Tet1) protein. This protein can bind to 5meC bases resulting in a conversion to 5hmC in both cultured cells and *in vivo* (Tahiliani, Koh et al. 2009). This protein may be catalysing the demethylation of cytosine nucleotides through a 5hmC intermediate however further work is required to determine whether or not this protein is indeed the elusive mammalian DNA demethylase.

CGIs as relics of early embryonic transcription

The fact that CGIs are often found at the 5' end of a gene may be an important factor in their origin and maintenance within a largely methylated genome. Early studies on CGIs first discussed the importance of transcription upon island methylation during early embryonic development. Particularly compelling evidence came from studies on the non-methylated CpG Island at the mouse adenine phosphoribosyltransferase (*aprt*) gene. In a transgenic experiment, binding sites for the transcription factor Sp1 were mutated at the *aprt* promoter resulting in a loss of transcription (Macleod, Charlton et al. 1994). Following pro-nuclear injection the resulting mice were seen to gain methylation over this CGI, presumably as a direct result of loss of transcription. Recent studies reveal that the majority of genes are indeed transcribed in ES cells, however only a subset of these genes produce viable full length transcripts. Subsequent ChIP-chip analysis (Chromatin immunoprecipitation followed by microarray detection) revealed that the initiating form of RNAP II is found at around 75% of promoters in ES cells (Guenther, Levine et al. 2007). The fact that similar numbers of ES cell promoters undergo transcription (be it abortive or full length) as contain CGIs indicates that early embryonic transcription may in some way protect these CGIs from the acquisition of methylation (Macleod, Ali et al. 1998).

Protection of CGIs through protein interactions

One of the simplest models for the origin and maintenance of non methylated CGIs is that they are bound by a group of proteins which results in protection from the Dnmts either directly or indirectly. To date studies have identified only a few candidate factors including the aforementioned transcription factor Sp1 (Braghetti, Piazzi et al. 1993). Even if CGIs are bound and protected by a specific group of proteins they are still highly accessible regions of the genome, as shown through DNA footprinting and nuclease sensitivity assays (Lin, Tomzynski et al. 2000).

Recent studies have found that the DNA methyltransferase cofactor Dnmt3L is not able to bind to histone H3 tails where the lysine 4 is di- or tri- methylated (Ooi, Qiu et al. 2007). *De novo* methylation during development may therefore be excluded from regions of the genome containing

high levels of this histone modification. As the promoter regions of active genes have previously been shown to correspond with high levels of the H3K4me3 modification (Bernstein, Kamal et al. 2005; Kim, Barrera et al. 2005; Roh, Cuddapah et al. 2006; Barski, Cuddapah et al. 2007) these loci may not be targeted for methylation by the *de novo* methyltransferases.

As CGIs are often found associated with the 5' ends of genes the transcriptional machinery itself may act as a protective shield against the methylation seen elsewhere in the genome. The initiating form of RNA polymerase is shown to be present at around 75% of promoters in ES cells (Guenther, Levine et al. 2007) and CGIs may therefore remain methyl free due to the direct protection afforded by this protein complex and not transcription *per se*. This model is also valid within the contexts of the studies on the *aprt* CGI, as abolition of Sp1 sites would lead to loss of any RNAP machinery based protection.

1.3.4 Methylation of CGIs

Although CGIs are usually defined as non-methylated clusters of CpG dinucleotides, recent work has found that in a minority of cases these islands can become methylated. The methylation of islands occurs both naturally and abnormally with the end result of gene silencing.

Dosage compensation in mammals is carried out through X chromosome inactivation to ensure that equal numbers of X linked genes are expressed in both sexes (Lyon 1961) . As females contain two copies of the X chromosome and males only one, lack of such a system would result in uneven levels of gene expression between males and females. To balance the expression levels of these genes, one of the X chromosomes in females is silenced (termed the Xi). Methylated CGIs tend to be associated with the promoters of silenced genes on this inactive X chromosome (Norris, Brockdorff et al. 1991), (Norris, Patel et al. 1994). The mechanisms by which methylation is targeted to such CGIs are unknown.

Another example of CGI methylation resulting in the control of gene expression is that of imprinting. Imprinting a phenomenon whereby allele specific gene expression occurs depending on the parent of

origin (McGrath and Solter 1984; Surani, Barton et al. 1984) and is known to affect around 40 genes. These include the much discussed *H19* and *Igf2* loci. Allele specific methylation of the CGIs at these two genes corresponds with the silencing of either the paternal (H19 silencing) or maternal (igf2 silencing) copy of the gene (Sasaki, Ishihara et al. 2000). As this imprinting mark must be erased and re-established depending on the sex of the individual this process is entirely epigenetic.

Recent work has revealed that the methylation of CGIs also has a role to play in the differentiation of cell types. A class of methylated CGIs was first discovered by comparing the methylation profiles of promoter regions of differentiated whole blood cells to those of testis and ES cells (Shen, Kondo et al. 2007). A subset of CpG island promoters (4%) were seen to gain methylation during differentiation and this was seen to correspond to the silencing of associated genes.

Finally, abnormal methylation of CGIs is widely reported in many cancers. Both hypomethylation and hypermethylation have been associated with the progression of cancers however the latter has been the most extensively reported. The majority of hypermethylated CGIs identified in cancers are associated with genes involved in cell cycle regulation (such as *p15* and *Rb*), DNA repair (*BRCA1* and *MGMT* – a p53 related factor) or apoptosis and proliferative control (*APC*) (Dobrovic and Simpfendorfer 1997; Stirzaker, Millar et al. 1997; Esteller, Garcia-Foncillas et al. 2000; Virmani, Rathi et al. 2001; Harden, Tokumaru et al. 2003). The methylation of these CGIs results in loss of expression of such genes and progression of a cancer state. As such the protection of CGIs from the methylation machinery is of great importance towards to progression of tumour formation.

1.4 The CXXC proteins

By analogy with the MBD domain which specifically targets proteins to methylated CpG dinucleotides, there exists a domain which instead targets proteins to non-methylated CpGs. This non-methyl CpG binding domain, more commonly referred to as a CXXC domain, forms a zinc finger structure which contains a series of highly conserved cysteine residues. This domain is found in a small group of proteins termed the CXXC family of proteins (Figure 1.4-1). Many of these CXXC

proteins have roles linked to the regulation of epigenetic state, including a DNA methyltransferase (Dnmt1), histone methyltransferases (Mll1 and Mll2), a histone demethylase (Kdm2a), a methyl binding protein (Mbd1), a candidate DNA demethylase (Tet1) as well as several proteins with as yet unknown functions. As these proteins all contain a CXXC zinc finger domain which specifically bind to non-methylated CpG dinucleotides, this family of proteins may act in a similar (albeit opposite) fashion to the MBD family and exert an effect on the local chromatin.

The CXXC domain itself binds to DNA through interactions with zinc cations and replacement of these with any other metal leads to loss of binding (Lee, Voo et al. 2001). Mutation of any of the conserved cysteine residues abolishes DNA binding and as yet no true consensus sequence has been identified other than for the presence of an unmethylated CpG dinucleotide (Lee, Voo et al. 2001). In depth investigations into the structure and DNA binding properties of the CXXC domain from Mll1 reveal that the domain makes up a crescent like structure incorporating two zinc ions (Figure 1.4-2, A). Each of these ions is contacted by three highly conserved cysteine residues (Figure 1.4-2, B) forming a loop like structure thought to interact with the underlying non-modified cytosine nucleotide. The overall DNA binding interface is made up of a groove of positively charged residues (Figure 1.4-2, C) (Allen, Grummitt et al. 2006). Mutation of any of these residues (R1154, K1176, K1178, K1186 or K1193) results in loss of DNA binding without the unfolding of the protein whilst mutation of any of the conserved cysteines results in complete unfolding of the protein (Allen, Grummitt et al. 2006). The structure appears to only be able to bind to non modified CpG dinucleotides as methylated CpGs may sterically hinder the binding of the extended loop around the zinc ion.

The specific affinity of the CXXC domain from MBD1 towards non-methylated CpG rich DNA has been used as a mechanism to isolate a CGI DNA library allowing thorough investigation of CGI distribution and properties (Illingworth, Kerr et al. 2008). This result however is unlikely to represent Mbd1 binding *in vivo* as the CXXC domain alone may not represent the true distribution of Mbd1. As the CXXC domains within several proteins are highly conserved (see chapter 5.2), specificity of binding to particular regions of the genome may arise through either the total structure of these proteins and/or through the binding of specific CXXC protein cofactors (such as transcription factors).

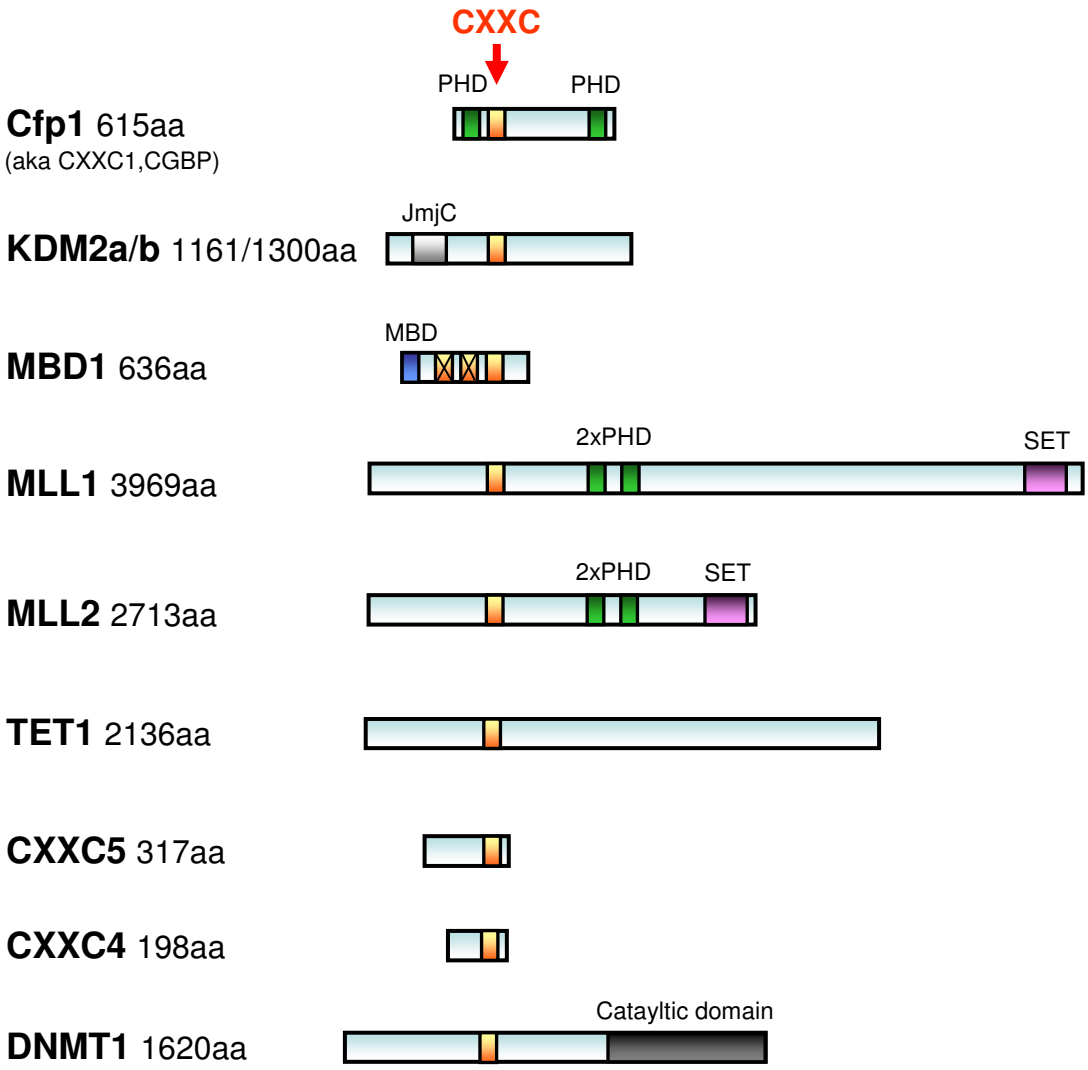

Figure 1.4-1. **The CXXC family of proteins.** The nine members of the CXXC protein family aligned through their conserved CXXC domain (yellow). These proteins convey a variety of epigenetic roles ranging from the modification of histone tails (Kdm2a, Mll1 and Mll2) to the methylation of DNA (Dnmt1). Several of these proteins including Cfp1, Mll1 and Mll2 contain chromatin binding PHD domains (green). Additionally Mll1 and Mll2 have histone methyltransferase SET domains (purple). Interestingly the Mbd1 protein, which has three CXXC domains (of which only one is functional – CXXC3) also contains a methyl binding domain (MBD – blue). The function of both CXXC4 and 5 are unknown to date.

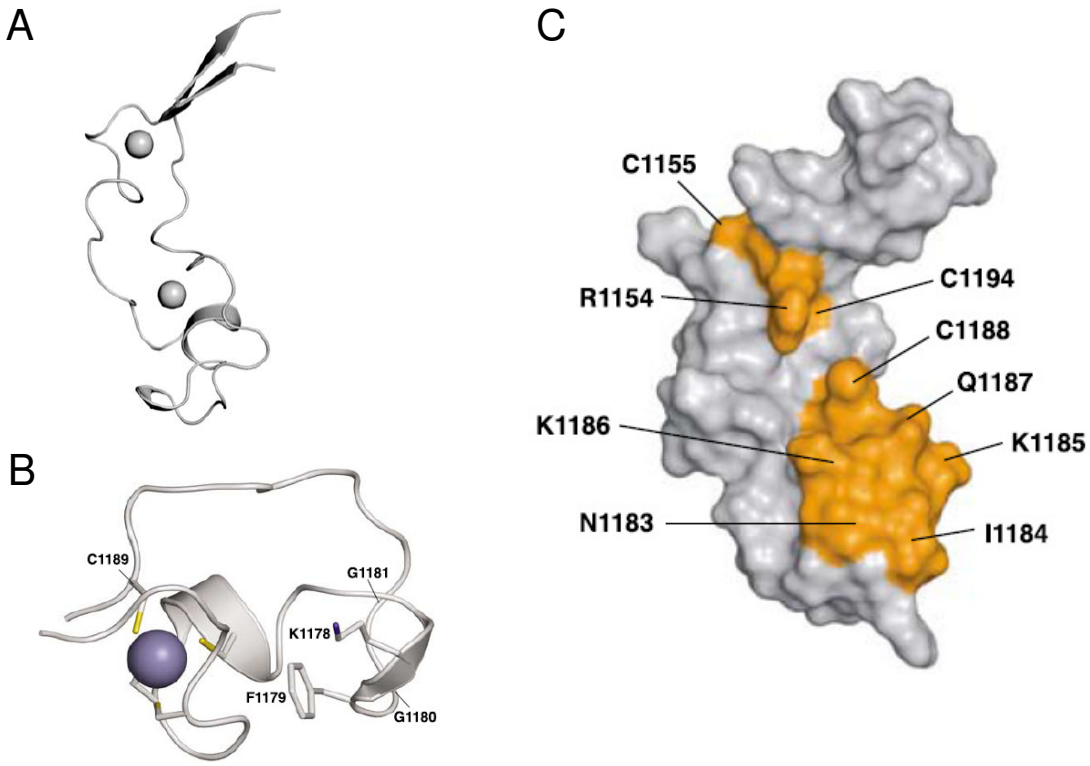

Figure 1.4-2. Structure of the CXXC domain from the Mll1 protein. **A.** The ribbon structure of the CXXC domain revealing the positions of the zinc atoms, represented as spheres, and the overall crescent shape of the domain. **B.** Close up of the extended loop (from 1181 to 1189) and second of the zinc atoms (blue sphere) contacted by cysteine residues (yellow bars). **C.** Representation of the molecular structure of the CXXC domain including the residues predicted to contact the DNA. Surfaces in yellow represent the contact surface between the protein and the DNA. (Figure edited from (Allen, Grummitt et al. 2006)

1.4.1 CpG binding protein, Cfp1

CpG binding protein 1 (Cfp1) was first identified in the year 2000 when the gene was cloned and found to contain two PHD domains as well as a CXXC, coiled coil and basic domains (Figure 1.4-1) (Voo, Carlone et al. 2000). Further studies found that this protein exclusively bound to non-methylated CpG rich oligonucleotides as well as acting as a transcriptional activator as shown by its ability to activate luciferase promoters upon cotransfection into cultured cell lines (Voo, Carlone et al. 2000). It has also been implicated as a component of the nuclear matrix and localises to euchromatic nuclear speckles (Lee and Skalnik 2002) again linking it to transcriptionally active chromatin. Subsequent work on this protein has identified it as a component of both Set1a and Set1b H3K4

HMTase complexes (Lee and Skalnik 2005; Lee, Tate et al. 2007) and there is some evidence also that it may bind to two other CXXC proteins; the H3K4 HMTases MLL1 and MLL2 (Ansari, Mishra et al. 2008). Interestingly Cfp1 is analogous to the yeast protein Spp1, a component of the yeast Set1 complex (Voo, Carlone et al. 2000; Miller, Krogan et al. 2001; Takahashi, Lee et al. 2009). Sequence alignments of the two proteins reveal high sequence conservation barring a few regions, most notably the absence of a CXXC domain in the yeast protein. Furthermore Spp1 and can bind to H3K4me3 modified histone tails via its PHD domain (Murton, Chin et al. 2010) and has been shown to interact with the origin recognition complex (ORC) through co-immunoprecipitation with the ORC2p protein (Kan, Zou et al. 2008). This Complex is implicated in many cellular activities including transcriptional control and the initiation of DNA replication foci.

Targeted disruption of the *Cfp1* gene leads to peri-implantation embryonic lethality in mice (Carlone and Skalnik 2001) which is the developmental stage during which the genome undergoes global loss and remodelling of cytosine methylation patterns (Reik, Dean et al. 2001; Li 2002). *Cfp1* -/- ES cells are viable although they exhibit an extended doubling time, largely due to apoptosis (Carlone, Lee et al. 2005). These *Cfp1* -/- ES cells fail to survive when stimulated to differentiate and therefore it is an essential factor in development. This may in part be due to a 60% reduction in the levels of cytosine methylation in these cells (Carlone, Lee et al. 2005). Although transcription of the major maintenance DNA methyltransferase enzyme (Dnmt1) is found to be reduced to around 50% of its wild type levels (Carlone, Lee et al. 2005) this alone cannot account for the dramatic phenotypes seen in the *Cfp1* null cells. *Dnmt1* -/- ES cells exhibit ~90% reduction in the expression of this protein with dramatic loss in DNA methylation levels, but are otherwise able to grow normally until differentiation (Lei, Oh et al. 1996). Thus the Dnmt1 reduction alone in *Cfp1* null ES cells cannot explain the growth defects seen.

Not only are the epigenetic modifications to the DNA altered upon Cfp1 removal, so to are the histone modifications, specifically that of H3K4 methylation. However, there is some disparity regarding the relationship between the CXXC protein and the active histone modification. In one study, *Cfp1* -/- murine ES cells exhibited a four fold increase in the levels of this active histone modification (Carlone, Lee et al. 2005). However this finding is somewhat contradictory to the results of a study

carried out in cultured human embryonic kidney cells (Ansari, Mishra et al. 2008). Levels of Cfp1 were depleted in these cells by using an antisense transcript resulting in the reduction of H3K4me3 signal from the *HoxA7* gene. Such a large disparity may represent discrete roles for the protein between the pluripotent ES cells and the differentiated somatic cells (Ansari, Mishra et al. 2008). Alternatively this may simply represent a difference between RNA depletion and knock out methods of analysis.

Alongside roles in the epigenetic control of both the DNA and histone tail modification, loss or removal of Cfp1 appears to result in the reduction of gene expression. The aforementioned study in human embryonic kidney cells found that transcription of the *Hox*A7 gene was reduced with a loss of both Cfp1 and H3K4me3 over the promoter region (Ansari, Mishra et al. 2008). Additionally, *Cfp1* -/- ES cells exhibit both a susceptibility to DNA damage as well as an associated loss in the expression of AP endonuclease (Ape1) activity, an enzyme involved in DNA base excision repair (Tate, Fishel et al. 2009). Interestingly, both the expression of this gene and the susceptibility towards DNA damage were returned to normal upon restoration of normal Cfp1 levels. Taken together, these findings reveal that the Cfp1 protein may be an important regulator of transcription. In the case of the *HoxA7* promoter region, the loss in transcription appears to occur due to both a loss of Cfp1 as well as H3K4me3.

1.4.2 Trithorax group proteins, MLL1/2

The MLL family of proteins are histone methyltransferases, acting on the fourth lysine of the histone H3 tail through their conserved SET domains. Like many of the histone modifying enzymes they exist in multi-protein complexes consisting of several shared subunits, and require these to exhibit their effect (Dou, Milne et al. 2006). Two of the family members, Mll1 and Mll2, also contain a highly conserved CXXC domain, similar to the one seen in Cfp1.

Mll1 (mixed-lineage leukaemia 1) is a 3969aa protein which has been shown to interact with components of the basal transcription machinery as well as human homologues of the yeast

COMPASS complex (Yokoyama, Wang et al. 2004). Chromosomal translocations that fuse MLL1 to other sites are often found in many myeloid and lymphoid leukaemias (Ayton and Cleary 2001) and there is evidence to suggest an important role in the regulation of the Hox genes during development (Ernst, Fisher et al. 2004). Most interestingly, ChIP-Chip studies have revealed binding of Mll1 over promoters at sites of high H3K4me3 and RNA pol II (Guenther, Jenner et al. 2005). These peaks were largest at the most highly expressed genes suggesting that MLL1 is possibly a binding factor of active genes, either through its binding to the transcriptional machinery, to the H3K4me3 mark itself or via its CXXC domain depending on the methyl status of the promoter.

There is some evidence that both MLL1 and MLL2 bind to the CXXC protein Cfp1 as flag immunoprecipitations co-purify MLL1, MLL2 and the hSet H3K4 HMTase proteins (Ansari, Mishra et al. 2008). Cfp1 may therefore interact with two or three independent H3K4 HMTase complexes.

1.4.3 Other CXXC proteins

A few of the other CXXC proteins also contain functions linked to the control of epigenetic modifications. The first of these is the lysine specific demethylase 2a (Kdm2a) which is responsible for the removal of methyl groups from histone H3K36 residues (Tsukada, Fang et al. 2006). As methylation of the H3K36 residues is important for the progression of transcription through the gene body, modulation of this mark must be tightly controlled. This control is imposed by Kdm2a by the removal of H3K36me3 marks from the body of genes once transcription has occurred, catalysed by Kdm2a's highly conserved JumonjiC (JmjC) domain (Klose, Kallin et al. 2006).

Another CXXC domain containing protein which has attracted much attention recently is the Ten eleven translocation 1 (Tet1) protein. Recent work has identified this protein as a potential demethylase through the conversion of 5-methyl cytosine into 5-hyroxymethylcytosine (Tahiliani, Koh et al. 2009). Although the functional significance of 5hmC is unknown it might be acting as either a signalling module for the recruitment of specific proteins (in a similar vein to methylated cytosine and the MBDs) or existing as an intermediate in the demethylation of 5mC (Kriaucionis and

Heintz 2009). As the domain architecture of the Tet1 protein is still largely unknown further roles for this protein have not been identified. The protein is known to contain a CXXC domain which leads to the possibility that this complex can bind to non-methylated CpG rich DNA such as that found at CGIs, which may in turn infer CGI maintenance roles upon Tet1.

1.5 Phd Aims

To date, CpG island research has largely focussed on the mapping these regions within the genome along with the study of the methylation states of these CGIs across various tissues and cells types (Antequera and Bird 1993; Antequera and Bird 1999; Shen, Kondo et al. 2007; Suzuki, Sato et al. 2007; Suzuki, Kerr et al. 2007; Illingworth, Kerr et al. 2008; Meissner, Mikkelsen et al. 2008; Straussman, Nejman et al. 2009; Illingworth, Gruenewald-Schneider et al. 2010). As CGIs are unique in their DNA sequence composition (significantly rich in CpG dinucleotides compared to the rest of the genome) and are known to associate with histone marks generally associated with an euchromatic state (Tazi and Bird 1990), CGIs may also contain a host of unique binding proteins.

The focus of the work in this thesis will be to attempt to identify any proteins which specifically associate with CGIs. This study will be expanded to contain both the histone tail modifications as well as any chromatin binding proteins. It is hoped that by identifying any uniquely bound protein factors that we may better understand: i) how these CGIs remain stably non methylated compared to the remainder of CpGs in the genome which are mostly methylated, ii) attempt to understand any possible functional roles for these CGIs and iii) to investigate the effects (if any) that the underlying DNA sequence at a CGI exerts towards defining the local chromatin environment through the binding specific protein factors.

Chapter 2: Materials and Methods

2.1 Materials

All reagents stored at room temperature unless otherwise stated

2.1.1 Nuclei and chromatin manipulation

Nuclei Homogenisation Buffer: 10mM Hepes pH7.9, 25mM KCl, 0.15mM Spermine, 0.5mM Spermidine, 1mM EDTA,2M Sucrose, 10% glycerol. Stored at -4°C.

Nuclei Storage Buffer: 50mM Tris-HCL pH8, 100mM NaCl, 5 mM MgCl2, 0.3mM EGTA, 10mM sodium butyrate, 1mM β-metacaptoethanol or DTT, protease inhibitors. Stored at -20°C

Nuclei Lysis buffer: 1% SDS, 10 mM EDTA, 50 mM Tris-HCl (pH 8.1), protease inhibitors. Stored at 4°C

2.1.2 Nucleic acid Manipulation

TE buffer pH7.5: 10mM Tris HCl p7.5, 1mM EDTA

TAE electrophoresis buffer (1x): 40mM Tris-acetate, 1mM EDTA

TBE electrophoresis buffer (1x): 45mM Tris-borate, 1mM EDTA

Blue DNA loading buffer (6x) 2.5mg/ml bromophenol blue, 0.3mg/ml glycerol, 120mM EDTA pH8.0, 4.2% (w/v) SDS. Stored at -20°C (long-term) or r/t (short-term).

Orange G loading buffer (6x): 0.198% (w/v) orange G, 12% (w/v) Ficoll, 120mM EDTA pH8.0, 4.2% (w/v) SDS. Stored at -20°C (long-term) or r/t (short-term).

DNA Sequencing Buffer (2.5x): 20mM Tris HCl (pH8) and 5mM MgCl2.

Bisulfite modification solution (pH of 5.0): 3.8g sodium hydrogen sulfite (NaHSO3) was dissolved in 5ml dH2O and 1.5ml 2M NaOH (kept away from light). 110mg hydroquinone was dissolved in 1ml dH2O at 55°C for ten minutes. The sodium bisulphite and the hydroquinone solutions were then mixed. Solution is prepared immediately before use.

RNA extraction: Tri- Reagent (Sigma-Aldrich)

Oligonucleotides: Custom Oligonucleotides were purchased from Sigma-Genosys. All were resuspended in dH20 to 100 µM whilst working stocks were diluted to 10µM. All resuspended Oligonucleotides were stored at -20°C.

Table 2.1-1. ChIP PCR Primers

ID	forward	reverse
actb		
1.2	CAGTACATAATTTACACAG	CCAAGTATCCATGAAATAA
1.3	AAGCCATGCCAATGTTGTC	AGCAGATGTGGATCAGCA
2.1	TGAAGCTGTAGCCACGCTC	CTGTCCCTGTATGCCTCTG
3	GCAGAAACTGCAAAGATCC	CCACACCTTCTACAATGAG
3.1	CCAAAGTAACAGGTCACTT	GTGTCTTGATAGTTCGCCA
4	CCTAATACGGCTTTTAACA	CCTGAGGATCACTCAGAAC
4.2	CGAGCACTTAAGTGGATGA	GCTTTCCGGCTATTGCTAG
4.3	CAGTGGTCCAATAACTGGA	CCAGGCTTAGTCTTGCCTG
5	CCTTGTCTGGAAGAGGTGA	CTGAGCAACTGAGAAATAC
6	CCTAAGCTGCACATTTTCA	GGCATTATGGCTGGATCTT
Atg9b		

1	GGCCTCTGTCTGTCCCTCT	AATGCCTATGCCTTGCTCC
2	GCGGAACCAGAAGAGCAGA	GCTTCGGAATGAGTGGTTG
3	GTCAGAGAAGGGGCACACA	CTTCTCGGGAGCTCTCTTT
4	CCCTTGTTGGCTAACGCTA	GTCCCTAGGCTTCCCTGTT
5	CAGCATCTTCCCCTCTGCT	CCATCCTTTGCCTTCCTCT

bdnf		
1	TTCCTGATCTATGAGCAGA	GACAGCTAAGAAATCTCTC
3	GTGCTCAACTCTGAAATTC	CCTTTCTTGATTTCTCCCT
4	GGAAAATATACACATGTGT	CCTAAGAATGCTTCTTTAA
4.6	GCACAGACAGTTCCGGGATCCAGG	GGAGTTCCTTTGCAGCACGCTTGG
5	CATTCTTAGGATAGGAGGG	CGAAAAGGTGTAGGCTAGG
6	CCTTTTCGCTCAGAGCTCC	GTGTGACTTATGAATCTAA
7	CGATCCACTGAGCAAAGCC	CTCCAAAATCTGACTCTCT
9	GCTTCATTGAGCCCAGGTC	GCACGCTTGGAATTGCATC
12	CTGATCTGTCAGTGCATTG	GAGCCTCCAACTTTGAGAG
14	GACAACAGATTTGGGACCT	GCACTGATATAGTCTCTGA
14.6	TGGAGAAGGAAACCGCCTGGGG	AAAGCTCTCGGATCCCCGGCAG
15	GGAAGTTAAGACAGAGATC	GCTCAATTGAAGCACATAT
17	CCATGACAATCTCTCTAGG	CCTCATGTGACTTTAGCAT

C-myc		
1	CCAGAAGCTTTCCCAGCAA	CAGCTCAGCCTTGCTTGCT
2	GCATGCCATGGCTAGCTTG	GAGAAGCAGGAGACCCTAA
3	AGTGACTGAATGAATGCTG	TCTCATACTAACAGTCATG
4	GTGATGTCATCAGGCTGGG	CACCAACAGATATGACAGT
5	AGGCAATAATAAGCTAATG	GTGAAGGAGATCTAACAGA
6	TGGTTAATAAGCTAGATTA	CCTTCGTATGTGTGTGTTA
7	CCTAGATAACTCATTCGTT	CCCTGCGTATATCAGTCAC
8	GCTGTTTGAAGGCTGGATT	CAACTACTCTTGAGAAAAG
9	CAGTGCTGAATCGCTGCAG	CCGATTGCTGACTTGGAGG
10	GGAAGAGAATTTCTATCAC	ACATAGGATGGAGAGCAGA
11	GGCTTATCTTTCAGCTCCA	TGGGTCTTAGACAAACGTA
12	CTCAACCCAAGGACTCTGC	CCAGGATCAACTTAGCAGT

Dlx5/6		
-2	TCCATGATCTCCTAGGTAT	TGAAAGCTATTCAAATGAA
3	CAAAGATGACACAGTCAAA	CTCACTCCTTTCCACCTCT
6	GGTTCTACCACCTGGATGT	CTCTAATTATACACCAGTG
7	GAGCTAAGGTGGCTGCAAA	GGATTTGGACGAGTCCTGG
8	CTTACAGCGCCACGGACTT	GCTATACCACTGTGGGCAC

9	GTAATGCTACATTGTAGGT	CTACCTCCTATGTTGCTTA
12	GAGACTACAGCTCTATGCA	GCAGTCAGCTATGGGGATG
16	GCATAGGCTAGATATCTAC	CTCTTTCAGGATCTCCTGT
18	CCATCCTCAGATCATACAT	GAAACACACAAGGTGAATA
19	GTGCTGTCCTGTTTGCAGT	CCCCTACCACCAGTACGGC
21	CCTGAAGTGCTGAAAGACT	ACTCGAGGCTGTCTATAGT
23	CTTTGCAAACCAAAATCGT	CACCTGACAGTTCCGTTCT

Foxd3		
1	GCCTAGTCGGTGAGGAGTTG	CAAGCCTGGGTAGTCGAAAG
2	TCTTACATCGCGCTCATCAC	TCTTGACGAAGCAGTCGTTG
3	CGCTGATGATGCAGAGCTT	CCCAGGGTGTTGAGCTGTAG
4	CCGAATTTCCAAAAATGCAC	AAACTGCGCAGAGTGAACCT
5	CTCCTCTCGAGCCTTTCAGA	AGAATTGATGCCGTTTGAGG
6	GAGCCCTGCTGAAGAAGCTA	GAAGCCCTGCCTTTCTCTTT
7	CTCCCGAGCTGTTTTTGAAG	AATTCGCCTCCAGACCCTAT
8	AGGATGGGTCACAGAACCAG	TCTTCCATTTCCCAACTTGC
9	GACCTGGAACGCATTAGGAA	TGGAGTGAGGGTGCTTCTTT
10	TGACAGGGGACTCTTGCTCT	GGAGAGGCGAGCTAACTGTG

intergenic ch5		
1	GAAGGCACACACATGCAAA	GCGGCTACTGTTAGAAGAA
2	GACCAATGTCCGCTTAGCT	GCCGTGGCTTAGTGAAGAG
3	GATCCTGACTAGTGCCCCT	GGATCGCTGGATTCAGACC
4	GGGGAGAGGGAACAGTGGT	TGACTGTGTCCTGAGTGTG
5	GGAGCTGCCTGCTACTCAT	CCAGTGGAATTAGAGGGGC

Oprd1		
1	TTTTACGGGAAGCCACTTTG	CCATGTTTTCCACCATGACA
2	CAGATAGCTGTGAGCCACCA	TGGATAAGAATGGCCCTCAT
3	AGATGGCTCAGCGGTTAAGA	GTTCATGAGCCTTCCCACTC
4	GGCTGGATTGGAACTGGTTA	TCTGGTTCTAGGGGCTCTGA
5	GAGCAGTCGGGTGCTCTTAC	GGATTTAGAGGTGGGGGAGA
6	AGTTTTCATCCGGAGCCTTT	AGGGTTTTGGGGAGATGAAG
7	TTTCTTCATCGTGCAGTCACA	GCTCCGCTGACTCCAGGT
8	GCCCCTCGTCAACCTCTC	GGACGATGCCAAACATGAC
9	CGCCCCAAGTCACATTTAGT	TATCCCGAGTTCCTGACCAC
10	TGGGAGCACTAAGACCCTGT	AGAGATCTGCTTGCCTTTGC
11	ATGCTGCAGTTGGCTGTATG	CTCTGATGATCCCGACCACT
12	AGTCGCCAAGGACACAGAGT	CAGCCTTGGCTACAGGAGAC

Sycp2		
1	CCAGCAGCTGTTCAGTGCA	TTCCTTGGTTGGCTCCTTT
2	GCAGTCCCCTGACAAAGCT	ACGGAGTCCCCAAGTCCCT
3	CAGCCGGAAGTGATGGAAA	CTGGGCATGAGGATTGCCT
4	AGCTCCAATGACAGCCCAA	AAGGGTGCCTCAGGGAGTG
5	CATGCATGCCAGACAAGTG	CACCCTCTAAAAGCCTGTC

Xist		
-3	GACAGCCTTATCCAGTGTC	GAACAGCGTAAAACTGTAA
-2	GTGGTCTCATTGGTTGGCA	GGAACATTTTATGTGGATA
1.21	GTCTTGAGGAGAATCTAGA	TTCTATACCAGTTCAGGCT
2	GAATTCAACAAGTAAGCAA	GCCAGAGTCATAGTGGATC
2.1	GCACTGTAAGAGACTATGA	CGCATGCTTGCAATTCTAA
3	CCTGTACGACCTAAATGTC	GTATTAGTGTGCGGTGTTG
3.1	CTCAGTTTAAGAGCAAAGT	GCTTGGTGGATGGAAATAT
4	AAAAAGTATGGAGGACATG	CGTGCAACGGCTTGCTCCA
4.1	AGGTCACACACCTGTCTAT	CCAAGGAGCCATTTTGTGA
5	GTCTCGTTGATTCACGCTG	GTTTATTCAGTCTGTGTGC
6	CCTCCATTCCTGTACACTT	CTTAATGTTAAGAATATGC
7	GCTACTGCTCATAGGTAGG	CATGATCTTTGGTAGATTG

Table 2.1-2. Genomic DNA PCR Primers

ID	forward	reverse
Oprd1	GCTTCTGGGCAACGTGCTCG	CCTCATAATCTCAGAGGACC
Ccne1	GGACACAGCTTCGGGTACGGG	GACCCGTCTCTCACAGCCAC

Table 2.1-3. Primers Used for expression analysis

ID	forward	reverse
Actb	CCAAAGTAACAGGTCACTT	GTGTCTTGATAGTTCGCCA
Bdnf	GTGCTCAACTCTGAAATTC	CCTTTCTTGATTTCTCCCT
Cfp1	CGAGAACGACAGTGATGACA	AGTGCTCAGGACACAGCACC
C-myc	TGGTTAATAAGCTAGATTA	CCTTCGTATGTGTGTGTTA
Gapdh	TACCCCCCAATGTGTCCGTCG	CCTGCTTCACCACCTTCTTG
Mll1	CACTTCCTCCATAGGCTCCA	ATGTTTAATCCGGGGTCCTC

Restriction digestion of DNA:

Table 2.1-4: NEB restriction enzymes

Enzyme name	Cut Sequence
AluI	AG|CT
HinPI	G/CGC
HpaII	C/CGG
MseI	T/TAA
MspI	C/CGG

2.1.3 Protein Manipulation

Proteinase K stock solution: 20mg/ml proteinase K (Sigma), 100mM EDTA pH 7.5, 2% (w/v) SDS. Stored at □20°C.

Antibodies: stored at -20°C

Table 2.1-5: Antibodies

Antibody [Applications]	Source/ Catalogue number	WB conditions
Primary Antibodies		
Anti-Histone H3	Rabbit polyclonal to Histone H3 (ab1791 – Abcam)	1:5000
Anti-Histone H3K4me3 [WB/ChIP]	Rabbit polyclonal to Histone H3K4me3 (ab8580 – Abcam)	1:1000
Anti-Histone H3K4me3 [ChIP]	Rabbit polyclonal to Histone H3K4me3 (07-473 – Upstate)	N/a
Anti-Histone H3K9me3 [WB/CHIP]	Rabbit polyclonal to Histone H3K9me3 (ab8898 – Abcam)	1:1000
Anti-Histone H3K27me3 [WB/CHIP]	Rabbit polyclonal to Histone H3K27me3 (07-449 – Millipore)	1:1000
Anti-Histone H3K36me3 [WB]	Rabbit polyclonal to Histone H3 K36me3 (ab9050 – Abcam)	1:1000
Anti-Histone H3 pan acetyl [WB/CHIP]	Rabbit polyclonal to acetyl Histone H3 (06-599 – Upstate)	1:1000
Anti-Histone H4K20me3 [WB]	Rabbit polyclonal to Histone H4K20me3 (ab9053 – Abcam)	1:1000

Anti-RNAP II [ChIP]	Mouse monoclonal to unphosphorylated RNA polymerase II (ab817 - Abcam)	N/a
Anti-Cfp1 Skalnik [WB] / ChIP]	Rabbit polyclonal to CFP1 (kind gift of D. Skalnik)	1:1000
Anti-Cfp1 H-120 [ChIP]	Rabbit polyclonal to CGBP: H-120 (Sc-25391 – Santa Cruz)	N/a
Anti-Kdm2a	Rabbit polyclonal to Kdm2a (ab31739 – Abcam)	1:500
Anti-Mll1	Rabbit polyclonal to Mll1 (3798-100 – BioVision)	1:250
Anti-Mbd1	Rabbit polyclonal to Mbd1 (ab3753 – Abcam)	N/a
Anti-Hdac1	Rabbit polyclonal to Hdac1 (ab7028 – Abcam)	1:2000

Seconday Antibodies		
peroxidase antimouse IgGs [WB]	GE Healthcare - NA934 WB	1:10,000 / 1:20,000
Peroxidase antimouse IgGs [WB]	GE Healthcare - NA931 WB	1:10,000 / 1:20,000
Infra-red anti-rabbit 680 IgGs [WB]	Licor – 32223	1:25,000
Infra-red anti-mouse 800 IgGs [WB]	Licor – 32212	1:25,000

2.1.4 Tissue culture and manipulation of Cells

HeLa & P-M- growth medium: Glasgow Minimum Essential Media (GMEM –Gibco), 1% Non Essential Amino Acids (NEAA –Gibco), 1% Sodium Butyrate (Gibco), 1% Pennicillin/Streptomycin (Gibco), 10% Fetal bovine Serum (FBS,Perbio)

NIH-3T3 Growth medium: Same as Hela. Transfected cell lines supplemented with 2µg/ml puromycin (Invitrogen) for selection..

Trypsin: TrypLE Express Stable Trypsin (Invitrogen)

shRNA Vectors for Cfp1 depletion: synthesised and cloned by Oligoengine. Vector backbone is called "pSuper puro" and contained the following shRNA target sequences:

Table 2.1-6: Sequences for Cfp1 shRNA depletion in pSuper vector

ID	Sequence
Target 986	5'-GAAGGUGAAGCACGUGAAG-3'
Target 1250	5'-CAGCCAACCGAAUCUAUGA-3'
Target 1920	5'-CUUCACCAAACGAUCCAAC-3'

Transfection reagents: Lipofectamine reagent (Invitrogen) and Opti-MEM reagent (Gibco)

2.2 Methods

All reactions were carried out at room temperature unless otherwise stated

2.2.1 Mouse Work

Extraction and storage of mouse organs

Mice brains and livers were harvested from a variety of different wild type mouse strains, mainly BALB/C and C57 strains. These mice were aged between 1 and 6 months and were of mixed gender (unless required for gender depend experiments such as that in section 5.3). Within particular experiments (such as ChIP-PCR runs), the age and gender of mice were kept constant. Harvested brains were instantly frozen in liquid nitrogen and then kept at -80 until needed.

Isolation of nuclei from mouse tissues

Nuclei were extracted by first grinding 8 brains or 4 livers up into a fine powder using a mortar and pestle along with liquid nitrogen. This powder was then transferred into a 30ml dounce containing 16ml of Nuclei buffer A. The mixture was then supplemented with fresh EDTA (final concentration of 5mM) and EGTA (final concentration of 0.5mM) along with protease inhibitors. The ground brain was then mechanically dounced on ice at (10 strokes). 9ml of the resulting homogenate was then layered onto a 3ml cushion of Nuclei buffer A in a plastic 13 ml centrifuge tube before centrifuging in pre-chilled SW40 rotor (13ml tubes) @ 72,700g for 40mins at 3°C.

The resulting pellet was then resuspended in 500µl of Nuclei storage buffer, prior to centrifugation at 425g at 4°C for 5 minutes. This pellet, the cleaned nuclei, is resuspended once more in 500µl of nuclei storage buffer and kept at -80°C until required. Typical mouse brains would generate around 0.4mg of DNA (~4^10^8 nuclei per brain).

2.2.2 Nuclei and chromatin Manipulation

Nuclei digestion

Harvested nuclei were resuspended in nuclei digestion buffer to facilitate their digestion.

Sucrose gradients

Sucrose gradients were constructed to separate out the chromatin fragments which resulted from the nuclei digestions. First a 60% weight/weight sucrose solution in "Buffer B" was made by heating and stirring 240g of sucrose powder in 160g of buffer B. This sucrose solution was the starting point for the generation of a range of sucrose concentrations. For a 5-30% sucrose gradient, 6 different sucrose concentrations were used and layered upon each other. Each of these stocks would weigh 10g and as such the following equation was used to determine how much of the 60% stock solution would be added:

% gradient required (as decimal) x final weight / stock sucrose % (as decimal)

1ml of each stock sucrose concentration was carefully layered on top of each other in a Beckman Pollyamer centrifugation tube capable of holding 13 mls. The sucrose gradients were then left for 1 hour on ice to allow the fractions to settle. A maximum of 0.8mg of brain DNA in a 500µl volume is loaded on top of the gradients before ultra-centrifugation at 4°C for 3 hours at 20,200g using a Beckman SW-40Ti swing out rotor (Beckman). 300µl Fractions were carefully collected from the resolved gradient using cut off blue tips and each fraction split into two eppendorfs in order to study both the DNA and protein compliments.

"DIGNAM" salt washing and methyl sensitive digestion of nuclei

Nuclei were resuspended in 250µl of nuclei digestion buffer prior to the dropwise addition of 250µl of 0.3M NaCl in digestion buffer. Washing the nuclei with this salt solution results in removal of any non specific or loosely bound factors from the chromatin into the supernatant. Centrifugation at 1,700g for 5 minutes pellets the chromatin leaving these non-specific proteins in the supernatant.

Pelleted material is resuspended in 500µl of fresh digestion buffer. Nuclei were then subjected to a two fold over digestion with either the restriction enzyme HinPI (GCGC) or MseI (TTAA) for 45 minutes at 37°C (both enzymes NEB). Centrifugation at 3,800g for 5 minutes retains the small digested fragments of chromatin within the supernatant. The DNA was then extracted from the supernatant by alcohol precipitation and the proteins by Trichloracetic Acid (TCA) precipitation.

2.2.3 Chromatin Immunoprecipitation and analysis

Protocol for Chromatin immunoprecipitation

Chromatin Immunoprecipitation (ChIP) is used to enrich for fragments of DNA bound by a protein of interest. Protocols vary slightly between ChIP carried out on nuclei from tissue and cells from culture. These nuclei or cells are first washed and resuspended in PBS and proteins fixed to the DNA by crosslinking for ten minutes at room temperature with 2% formaldehyde. This reaction is quenched by addition of 125mM glycine before washing through PBS. Samples are then resuspended in 450µl cold lysis buffer for 10 minutes to release the DNA from the nuclei. 150µl of this solution is aliquotted into separate eppendorfs prior to the addition of 1350µl of dilution buffer along with 3mM $CaCl_2$, 10mM $MgCl_2$ and protease inhibitors. Nuclei from tissues such as the brain (not cultured cells) are then subjected to mild levels of micrococcal nuclease digestion in order to aid later sonication steps. As such, 12 units of micrococcal-nuclease are then added for 20 minutes at room temperature and this enzyme inactivated through addition of 10mM EDTA and 5mM EGTA. Both cell and nuclei samples are then subjected to sonication so as to shear the chromatin, generating fragments with an average size of 500bp. Conditions for sonication vary depending on the concentration and volumes used. A typical reaction would be carried out on a volume of 1ml for 2 minutes at 25% power at 5 second intervals with each blast lasting 2.5 seconds. These fragments of chromatin were then added to blocked protein A sepharose beads (GE healthcare), blocked for an hour with 1mg/ml BSA and tRNA before washing through and resuspending in dilution buffer. This step is crucial in order to preclear the beads of residual chromatin binding. Like samples are then pooled and spun down gently at 400g for 15-20 seconds to pellet the beads and chromatin removed. Antibodies are then bound to this

material overnight at 4°C. Typical antibody concentrations used were between 1-2mg/ml per 25ug of chromatin. For a complete list of antibodies and working conditions used during ChIP see the materials section. 10% of the sample volume is kept back as an "input" control and does not undergo immunoprecipitation to allow for the comparison of the enrichment between IP'd material and non IP "genomic" enrichments.

The following day the precleared beads (see above) are added to the IP samples and left at 4°C for 2 hours. These beads bind specifically to the IgG chains of antibodies bound to the chromatin proteins of interest. Bead:chromatin complexes are spun down at 2,600g for 1 minute and undergo a series of washes followed by 2,600g spins as follows: 1xTSE1 buffer wash, 4x TSE2 buffer wash, 1x buffer 3 wash and 3x TE wash. Beads are finally resuspended in 200µl of extraction buffer and placed at 55°C for 55 minutes before shaking at room temperature for a further 20 minutes. Chromatin fragments previously bound to the beads should now be in the supernatant and beads can be removed through a fast spin at 15,700g for 2 minutes. Chromatin samples are spun through QIAshredder columns (Qiagen) at 15,700g for 1 minute to remove any left over sepharose beads. Finally 300mM NaCl is added to all samples, including the inputs and the samples placed at 65°C overnight to remove any crosslinks.

The final steps of the ChIP protocol include proteinase K (Sigma) and RNAseA (Sigma) treatment at 55°C for an hour before carrying out DNA precipitation and resuspension into 0.1M TE.

Typically 4 mouse brains would be split for 5xIPs and Input material. Each IP would be resuspended in 1150µl of 0.1 TE (enough for 108 ChIP PCR reactions for each IP which equated to 9 runs with 12 primer pairs).

Analysis of ChIP qPCR results

DNA fragments prepared by ChIP are tested for amplification across regions of the genome (see "DNA manipulation" regarding PCR). Regions of high amplification correspond to high levels of protein binding. Relative levels of binding are calculated against input, non IP material. iCycler iQ

software (BioRad) assigns a baseline fluorescence measurement based on the standard deviations calculated for cycles 2-10. This baseline bisects the fluorescence curves from each PCR in the linear amplification phase and indicates the cycle threshold value for that reaction (Lander, Linton et al.). An arbitrary measure of DNA quantity (Q) can be calculated using: $Q = 2^{-Ct}$. Averages of the two readings for each primer pair are then calculated and compared to the input. ChIP profiles are representative of those from at least 3 independent PCR runs using separate ChIP experiments. Often the relative levels of enrichment would vary quite significantly between IPs however, the overall profile generated would be nearly identical when plotted normalised to the average values over the region.

ChIP-Seq for genome wide sequencing

Immunoprecipitations are performed from a single mouse brain as described above prior to end repair by incubation at 20°C for 30 min with 3 U of T4 DNA Polymerase (NEB), 10 U of Polynucleotide Kinase (NEB), 2 U DNA Polymerase I Large (Klenow) fragment (NEB), 1x T4 DNA ligase reaction buffer (NEB) and 400 nM dNTPs. The enzymes were then heat inactivated at 75 °C for 20 min, after which the DNA was ethanol precipitated. A string of adenine bases were added to the 3' ends of the DNA by incubation with 5 U Klenow Endonuclease (NEB), 200nM dATP and 1x buffer 2 (NEB) at 37°C for 30 min. After heat inactivation of the enzymes, Ilumina paired end adaptors were then ligated to the processed ChIP DNA by incubation with 300 U of T4 DNA ligase (NEB), 1x T4 DNA ligase buffer (NEB), 7.5% PEG-6000 and 2 pmol of annealed Ilumina adaptors for 3 h at room temperature. Resulting ligated DNA was purified using MinElute PCR columns (Qiagen) and eluted in 10 µl water.

Bioinformatic analysis of ChIP-Seq data

Data files representing the genome wide binding sites for proteins of interest were received as processed files in which the DNA sequences had been mapped to genomic coordinates (known as a "wiggle" file). This data can be visualised against the genomic sequence using the Integrated Genome Browser ("IGB"; Affymetrix. Found at http://209.135.50.217/partners_programs /programs/developer/tools/download_igb.affx). The height of these reads indicates the depth of

sequencing, an increase of which raises the likelihood that the sequence amplified is indeed specifically enriched during ChIP. In order to attain meaningful statistics from this raw file a set of in house tools called were created by Alastair Kerr at the University of Edinburgh which can be found at http://bifx3.bio.ed.ac.uk:8080.

Initial raw wiggle files tend to contain high levels of background noise likely due to non-specific binding events during sequencing. As such this noise was removed by normalising the signal across the entire data set. This normalisation process takes into account the height of a user defined acceptable background level (the "depth of sequence"), the minimum run of sequence below the stated depth required to maintain read data during background subtraction (the "length"), along with the maximum gap allowed between two sequencing reads (the "gap"). For the criteria used for each ChIP-Seq experiment refer to table 2.2-1 below.

Table 2.2-1. Parameters used for peak finding calculations

	Background Subtraction			Peak finding		
Sequence Dataset	Depth of sequence	length	gap	Depth of sequence	length	gap
Cfp1	2	90	20	4	45	600
RNApolII	2	90	20	4	90	250
H3K4me3	2	90	20	3	90	700
H3K27me3	Published in Meissner et al., 2008.					

Each individual ChIP-Seq data set is processed first by background subtraction and then by peak finding to identify significant peaks. Background subtraction "length" refers to the minimum run of sequence (in base pairs) below the stated depth required to maintain read data during background subtraction. Background subtraction "gap" is the allowed gap within this run of sequence. The peak finding "length" corresponds to the minimum run of sequence at the stated "depth" required for identification as an enriched genomic locus. Finally for the peak finding "gap" value, individual peaks that are closer together than this value are knitted together as a single locus.

To generate a file capable to use in statistical analysis the normalised wiggle file must be converted into a "peak file" in which regions containing a high sequencing depth over a region of the genome

are represented as a box. The criteria to determine a genuine peak is entirely user based and as such is open to some interpretation. Genuine peaks are found by selecting on the basis of i) the length in base pairs in which the peak in sequencing occurs over, ii) the height or depth of this region and iii) the length of the gap that separates sequencing peaks. These values vary depending on the antibody used (Table 2.2-1). Resulting peaks can then be counted to determine total protein binding sites or compared for overlap with other peak finding data sets to determine co-localisation between proteins.

Access to ChIP-Seq data sets

High throughput sequencing data for Cfp1, H3K4me3 and RNAPII has been deposited in the Gene Expression Omnibus (GEO) under the accession number GSE18578.

H3K27me3 data sets published by Mikkelsen et al, 2007 were found in GEO under the accession number GSE12241

The genome wide CGI data set is available in the "John Thomson Sep 2010" folder within the "public" folder at http://bifx3.bio.ed.ac.uk:8080/library

2.2.4 DNA Manipulation

DNA extraction and precipitation

DNA was extracted through mixing with 1 volume of phenol:chloroform:IAA before precipitation by either 1 part isopropanol at r/t or 3 parts 96% ethanol (EtOH) at -80°C along with 1/10th volume of NaOAC. After centrifugation at 15,700g for 15 minutes pellets were washed with 70% EtOH before allowing to air dry in a fume hood for 5 minutes. Once all residual ethanol has evaporated the DNA pellets were resuspended in 0.1M TE and stored at -20°C.

Restriction enzyme digestion

Digestion of DNA through the use of restriction enzymes was carried out as per instruction of the manufacturer. Typically a range of 2-10 units of enzyme were used to digest 1µg of DNA at 37°C diluted in the appropriate buffer. Generally these reactions were carried out over the period of 45 minutes to an hour. Restriction enzyme recognition sequences are outlined in the materials section.

Gel electrophoresis

DNA was resolved by electrophoresis through the use of Sub-Cell system (Bio-Rad). Depending on the size of DNA fragments to be resolved, gels were made spanning 0.8-2% (w/v) agarose in either TAE or TBE (most often used if fragments were <300bp). Gels contained 0.5µg/ml of ethidium bromide which intercalates with the DNA and allows fluorescent visualisation upon excitation with Ultraviolet (UV) light. DNA samples and a DNA ladder of known fragment sizes were mixed with either 6x blue loading buffer or orange G loading buffer. Gels were run in the same buffer that they were made from and were run at a constant voltage no higher than 80V. Visualisation of DNA fragments was carried out over a UV light and images taken if needed.

Gel Extraction

Where DNA fragments were further required to be purified, single bands were cut out of the gel and placed into an eppendorf tube. Extraction of DNA from the gel piece was then carried out using the Perfect Gel cleanup Kit (Eppendorf) following the manufacturers' instructions.

Measurement of DNA concentration

DNA solutions were measured at OD260nm and OD280nm using a Nanodrop-1000 spectrophotometer (Thermo Scientific). The purity of the DNA was determined using OD260nm:OD280nm ratio, with a value greater than 1.8 indicating the absence of protein or phenol contaminants from the sample. An automated read out of the DNA concentration was determined through the use of Beer's law [Concentration in ng/ul = (Absorbance OD260nm x Extinction coefficient dsDNA 50ng/µl/cm) / pathlength cm].

End labelling of DNA fragments with radioactive dNTPs

1µg of DNA was typically labelled with 10µCi of radioactive dCTP (GE healthcare). Labelling was carried out by using the Klenow enzyme (Roche) which has 5' to 3' polymerase activity as well as some 3' to 5' exonuclease activity. To the DNA and radioactive dCTP mix was added 100mM dNTPs lacking dCTP as well as 1-5 units of Klenow enzyme. Reactions were carried out in the high salt buffer supplied (Roche).

Standard Polymerase Chain Reaction (PCR)

PCR reactions were used wherever it was required to amplify up DNA of interest from a DNA template. Reactions vary however typically were of a final volume of 25µl. In this reaction was the DNA template (1-100ng), 200-400µM dNTPs, Red Hot PCR buffer (ABgene), 5U Red hot DNA polymerase (ABgene) and 2.5mM $MgCl_2$. If amplifying up GC rich regions, PCR reactions were supplemented with 3% DMSO and sometimes slight increases in $MgCl_2$. Positive and negative controls were included by using a known amount of genomic DNA for the positive and dH_20 for the negative control.

For every set of primers, the number of cycles and annealing temperature (T^{an}) were different. Typical reactions however would compromise of an initial denaturation step for 2 minutes at 94°C followed by 30-40 cycles of denaturation at 94°C for 50 sec, primer annealing at T^{an} for 50 sec, primer extension at 72°C for 50 sec with a final extension phase after the cycles had been complete, at 72°C for 2 mins.

Quantitative PCR (qPCR)

This procedure was used whenever the need for quantitative comparisons of relative DNA arose. The two main procedures requiring qPCR were expression analysis of the same gene between cell lines and in the profiling of binding proteins to DNA regions after ChIP purification. The principles of qPCR are the same as standard PCR but for the incorporation of a dye in the reaction. The dye used, called SYBR green, is a cyanine dye that absorbs light at a λmax of 488 nm and emits light at a λmax of 522nm when complexed with double stranded (ds) DNA. Measurement of emitted light (522nm) directly allows the real-time quantification of duplexed DNA as it is synthesised. 5-50ng of DNA was

typically used per reaction and a specialised master mix (Quantace) containing SYBR green, dNTPs and $MgCl_2$ added. For expression analysis, one set of primers, typically designed to span an intron (thus remove DNA contamination, see RNA manipulation section for more), are then added to the reaction. Reactions are then carried out with a denaturation step at 96°Cfor 10 minutes, followed by 45 cycles of denaturation at 96°Cfor 50 secs, primer annealing at T^{an} for 50 secs, primer extension at 72°C for 50 secs with a final extension phase after the cycles had been complete, at 72°C for 2 mins. Ct values (see section on "analysis of ChIP qPCR results" above) are extracted from the post run data and compared between the two samples to determine relative expression levels. Internal controls are imposed to normalise the samples. This is usually accomplished through first normalising the Ct values to those of a stable housekeeping gene such as GAPDH.

In the case of ChIP qPCR many of the same principles apply with regards to the PCR reaction mixture and program of amplification. 5 to 12 primers were designed over regions of interest. Amplicons were not to be shorter than 150 or exceed 350 base pairs in length to give optimal amplification. On each 96 well plate profiles containing a maximum of 12 primer pairs were run in duplicate to limit rogue results. Each plate contained two rows of input (non IP) DNA. This would allow direct comparisons between the Ct values generated in the IP sample over a particular primer pair to no IP material, thus allow the calculation of an enrichment value of input. Each plate would typically contain 2 rows of 12 primer pairs for input material, with the remaining 6 rows of 12 divided between 3 IP samples, or 2 samples and a negative control, each run in duplicate. For more see the section on the analysis of ChIP results below.

2.2.5 Protein Manipulation

Whole cell protein extracts

Whole cell protein extracts were used to investigate total cellular levels of proteins. Typically 3-8^10^6 cells are washed in PBS before removal from culture dishes through the addition of 2mls of recombinant trypsin (Invitrogen). Once cells have detached from the culture plates the reaction is stopped by the addition of 1ml of culture medium before centrifugation at 400g to pellet the cells. Cells are washed in PBS including protease inhibitors and pelleted before resuspension in 50-100µl of

SDS loading buffer (without bromophenol blue dye). This solution is boiled for 5 minutes and 1µl taken for quantification by Bradford assay. At this stage bromophenol blue can be added to the protein solution in order to visualise on SDS polyacrylamide gels.

Trichloracetic Acid (TCA) precipitation of proteins

Certain experiments, such as the collection of sucrose gradient fractions, led to the attainment of large volumes of protein samples at low concentrations. In order to analyse these protein samples they first had to be concentrated to detectable levels. The most common way of concentrating protein samples is through TCA precipitation. 1% SDS is first added to the samples before adding TCA to a final concentration of 20% and incubating at -20°C o/n. The following day samples were thawed on ice before cold centrifugation at 15,700g for 15 minutes. The resulting pelleted proteins are near impossible to visualise in the eppendorf at this stage however the supernatant is carefully removed before adding 500µl of ice cold acetone and incubating on ice for 15 minutes. The protein pellet is encouraged to disperse in the acetone through gentle flicking of the eppendorf before the material is repelleted by cold centrifugation for 15 minutes at 15,700g. Once more the supernatant is carefully removed and the protein pellet left to air dry for five minutes on ice. Once all traces of the TCA and acetone have evaporated the pellet is resuspended in SDS loading buffer, boiled for 5 minutes to denature proteins and either used or stored at -80°C until required.

Bradford assay for protein quantification

Bradford assays were used to determine the protein concentrations of samples prepared, following the protocols outlined by the manufacturer (Biorad). BSA standards were prepared over a range of concentrations encompassing those expected in the protein samples. Typical final BSA concentrations in diluted Bradford reagent would range from 0.1 to 10µg/ml. Bradford reagent (Biorad) is diluted 1:5 in deionised water and added to the BSA standards and the samples in question to a final volume of 1ml. 1µl of the protein sample to be tested is taken for quantification and mixed with the diluted Bradford assay. This mixture will turn blue depending on the concentration of protein and the intensity of this colour is quantified by absorption at 595nm in a spectrophotometer. The

concentration of the protein sample can then be calculated through comparisons to a standard curve using the absorbances of the known BSA standards.

Protein Electrophoresis

SDS polyacrylamide gel electrophoresis (SDS PAGE) was used to resolve proteins based on molecular weight using the Mini-PROTEAN 3 system (Bio-Rad). Stacking gels contained 5% (w/v) acrylamide and separating gels between 8 to 15% (w/v) acrylamide depending on the molecular weight of the proteins of interest. Protein samples, prepared in protein loading buffer, and Pre-stained broad range protein markers (11-170kD; NEB) were incubated at 100°C for 5min to denature proteins and disrupt any higher order structures. Samples were then chilled on ice and loaded into the wells of the stacking gel. Electrophoresis was carried out at 25mA (Mini-PROTEAN 3 system, Bio-Rad) in Tris-glycine electrophoresis buffer.

Staining of gels to visualise proteins

SDS PAGE gels were stained by using either Coomassie reagent for a quick check of protein content or Silver stained for a more thorough analysis. For coomassie staining, gels were incubated in Coomassie Brilliant Blue R-250 staining solution for 20 minutes with gentle agitation. Gels were de-stained by rinsing several times in dH2O at 100°C resulting in the visualisation of all proteins as blue bands. The gels can then be scanned onto the LI-COR scanner (LI-COR) which not only creates an image file of the gel but allows for accurate quantification of specific bands as well as total protein levels. As such this approach can be useful in the quantification of protein samples to ensure equal loading on subsequent gels.

If protein levels are too low for the use of Coomassie staining a more sensitive approach can be used such a silver staining. This approach is based off the rationale that proteins bind silver ions, which can be reduced under appropriate conditions to generate a stained gel. Protein samples are run on SDS PAGE gels as explained above prior to soaking in 50% and then 5% methanol for 10 minutes respectively. Gels are then washed in 125ml dH_2O supplemented with 4µl 1M DTT for at least ten minutes. The gel can then be stained through a series of washes, firstly through a 10 minute incubation

in “silver solution” containing 0.2M silver nitrate, before washing off 3 times in dH_2O. To visualise the protein bands on the gel, developer solution containing sodium carbonate is added until the desired intensity is achieved and the reaction quenched by adding 5% acetic acid to the solution. The gel was then washed several times in dH_2O and visualised through scanning onto a computer or dried onto Wattman paper for archiving.

Immunoblotting

Once resolved on an SDS polyacrylamide gel, proteins were then transferred from the SDS PAGE gel to PVDF (Bio-Rad) or nitrocellulose membrane (Bio-Rad) using the Trans-blot SD semi-dry transfer cell (Bio-Rad) with constant voltage of 2mA for 70 minutes. PVDF was typically used for histone proteins due to its smaller pore size which do not allowing histones to pass through upon transfer. PVDF membranes are first blocked in TBS supplemented with milk powder, varying from 4 to 8% depending on the antibody in use. Membranes were blocked for 2 hours at room temperature before washing in TBS 3 times. Antibodies specific for the protein of interest were then diluted to their optimum working concentrations in blocking solution. Conditions for western blot are highly sensitive and vary between antibodies as well as between batches of the same antibody. Typical primary antibody concentrations ranged from 1:1000 for Skalnik-Cfp1 down to 1:5000 for Histone H3 (For the full list of antibodies and working concentrations see “Materials” section above). Primary antibodies were incubated on the membrane overnight at 4°C. The following day membranes were washed in TBS for 15 minutes before incubating for 2 hours at r/t with secondary antibodies specific for the organism the primary antibody was raised in. The majority of secondary antibodies were raised in rabbit or mouse. Typical concentrations in blocking solution range from 1:1000 for HDAC1 to 1:5000 for Histone H3 (for the full list of antibodies used as well as working concentrations see “Materials” section above). Membranes were then washed three times for 15 minutes in TBS to remove all of the secondary solution. The secondary antibodies bound to the membrane are conjugated with horseradish peroxidise and as such emit a signal when activated by enhanced chemiluminescence (ECL). ECL solution was washed over the membranes for 30 seconds prior to detection on light sensitive film. Resulting films were then scanned into a computer and bands corresponding to proteins of interest analysed for relative quantity or abundance.

LI-COR fluorescence blotting

The use of fluorescent secondary antibodies instead of the traditional horseradish peroxidase approach allows far greater sensitivity to be attained in a fully quantitative manner. Traditional western blotting approaches are difficult to quantify as saturation of resulting bands on the membrane occurs rapidly. Using the LI-COR scanner system in conjunction with fluorescent secondary antibodies can overcome this problem. However, this approach if not carried out efficiently can lead to high levels of background signal. Fluorencent blot protocols are the same as those for the western blot however the secondary used is fluorescent (typically at 680nm for red signals and 800 for the green). As this fluorescence is light sensitive, steps following the addition of this antibody are carried out in the dark. For a list of antibodies used and working conditions for each see the materials section).

2.2.6 RNA Manipulation

RNA extraction

RNA was extracted using Tri reagent (Sigma-aldrich). Tri reagent contains phenol and guanidine isothiocyanate and serves to maintain RNA integrity during cell disruption and dissolution of cell components. Subsequent centrifugation with chloroform allows biphasic separation such that RNA may be retrieved from the aqueous phase whilst DNA and protein remains in the organic phase. Typically 1ml of Tri reagent (Sigma-aldrich) was added to a confluent 100mm dish of cells in culture leading to their dissociation. 200µl chloroform was added, mixed vigorously and incubated for 10min at R/T. The aqueous phase (RNA) was resolved from the organic phase (DNA and proteins) by centrifugation for 15min at 12,000g at 4°C. To precipitate the RNA the aqueous phase was resuspended in 500µl of isopropanol, mixed and left to stand for 10min at r/t. RNA was pelleted by centrifugation (10min at 12000g at 4°C), washed with 70% EtOH, air dried and resuspended in 50µl of nuclease free water (Ambion). RNA was resolved by RNA gel electrophoresis (see RNA electrophoresis below) to determine the quality of the preparation as indicated by sharp bands representing the 28S and 18S ribosomal RNA. RNA was stored until needed at -20°C.

cDNA synthesis

To achieve accurate amplification of cDNA, contaminating DNA must be removed from the RNA prior to cDNA synthesis. Typically around 500ng of RNA was used in a total volume of 40µl containing 3ul (6U) DNAse I (DNA-free kit; Ambion) in DNAse I reaction buffer (DNA-free kit; Ambion) and incubated for 30min at 37°C. DNase I was inactivated according to the manufacturer's instructions. Primers for cDNA amplification were designed to span exonic sequences to reduce amplification of contaminating genomic DNA. These exons will only be close enough to amplify by PCR once processed into mRNA thus genomic products should be minimal. Prior to reverse transcription the RNA was denatured at 65°C for 5min and snap chilled on ice. 250ng of the RNA was prepared in a total volume of 25µl containing 5µM random hexamers (Roche), 1mM dNTPs (Abgene), 40U RNasin (RNase inhibitor; Promega), 200U of M-MLV reverse transcriptase (RNase H minus; Promega) and MLV reverse transcriptase reaction buffer (Promega). In addition a minus (-)RT reaction was set up in parallel containing all the components except for the reverse transcriptase to control for DNA contamination by PCR. Typical reactions would consist of 4 cycles each for 8min at 20°C, 8min at 25°C and 30min at 37°C followed by final heat inactivation at 70°C for 15min. The 25µl RT reactions were stored at −20°C. 1µl of each cDNA was used in each subsequent PCR reaction.

RNA Electrophoresis

RNA samples and size marker (281bp-6.58kb; Promega) were denatured in RNA loading buffer for 10min at 70°C and chilled on ice. These were then loaded onto a 1.5% (w/v) non denaturing agarose gel containing 0.5µg/ml ethidium bromide. Agarose gels were run at constant voltage (80V) in TAE buffer and visualised under UV light.

2.2.7 Tissue culture and manipulation of cells

General culture of cells

HeLa, NIH-3T3 and P-M- fibroblast cell lines were each cultured in their own growth medium as described in the "Materials" section. Cells were grown to confluency before washing in PBS and then splitting into new plates using recombinant trypsin (Invitrogen) to dissociate the cells from the plate. Cells are stored at –180°C stored in growth medium supplemented with 20% DMSO and an additional 10% FBS (Perbio).

Transfection of cells with shRNA

NIH3T3 cells were transfected using lipofectamine reagent (Invitrogen) with three independent pSuper vectors containing short hairpin constructs directed against *Cfp1* (Oligoengine) or vector alone. The target sequences for these sequences are described in the "materials" section above. Additionally transfections were carried out using a mixture of all three constructs. 20µg of vector DNA was mixed to a final volume of 800µl in Opti-Mem (Gibco) in an eppendorf whilst 45µl of lipofectamine reagent was taken to 800µl Opti-Mem in a second eppendorf. These two solutions were mixed after 5 minutes at 37°C and diluted to 8mls in Opti-Mem. This mixture was left at 37°c for a further 45 minutes to allow the lipid rich lipofectamine and the vector DNA to mix. A 90mm dish of NIH 3T3 cells at approximately 70% confluency were harvested and growth media removed prior to the addition of the mixed transfection solution. This solution was left on for 5 hours to allow transfection to take place. After 5 hours this solution was removed and replaced with standard growth medium once again. The following day these cells were placed under puromycin selection in order to kill all non-transfectant cells. Upon transfection and subsequent selection the majority of cells died leaving single cells to grow, forming colonies which were expanced for further analysis. This approach was carried out in triplicate with three vectors each containing a shRNA sequence designed to reduce the levels of Cfp1. Furthermore this shRNA transfection was repeated with a mixture of all three vectors in to attempt to create a superior knockdown effect.

Resulting sh*Cfp1* cells were cultured in 3T3 growth medium supplemented with puromycin at 2µg/µl with the minimum number of passages possible to avoid reversion of knockdown state.

Calculation of *growth* curves for *shCfp1* cell lines

Both control transfected ("pSuper clone 10") and cfp1 deficient ("shMix clone 3") 3T3 cell lines were seeded at $0.2x10^6$ cells each in 6 x 35mm tissue culture dishes. These cells were grown in GMEM supplemented with puromycin. Plates were trypsinised at selected time points (15, 19, 24, 48, 72 and 90 hours) and the number of cells calculated using a haemocytometer. Dead cells were distinguished from live cells by staining with Trypan-Blue (Invitrogen) and removed from the analysis. Cell numbers were plotted against the respective time point in order to determine a growth curve over the first 90 hours.

Chapter 3: Isolation and purification of CGI chromatin

3.1 Introduction

Eukaryotic DNA is packaged into a higher order chromatin structure compromised of repeating protein structures called nucleosomes. Each nucleosome consists of ~146 base pairs of DNA wrapped around an octamer consisting of four core histone proteins (two of each histone: H2a, H2b, H3 and H4) which in turn can be modified on their N terminal tails by a multitude of modifying factors (Kornberg 1974). These histone tail modifications can affect the local structure of the chromatin resulting in either compact or accessible regions. Modifications of these histone tails can affect the accessibility of the genetic code through the regulation of chromatin compaction as well as through the recruitment or repulsion of enzymatic protein complexes. Particular regions of the chromosome tend to be associated with a certain chromatin state. For example, centromeric sequences tend to be found in a heterochromatic state (Pluta, Mackay et al. 1995) whilst certain regions of the genome such as the promoters of active genes have been shown to correspond with particular sets of histone modifications permissive to transcription (Barski, Cuddapah et al. 2007).

Advances in techniques of chromatin purification along with mass spectrometry have allowed the sophisticated analysis of histone modifications to be carried out on a genome wide level. Further to this, recent advances in genome wide sequencing technology such as ChIP-Seq (see chapter 4.4) have lead to the high resolution mapping of histone modifications in a variety of cell types (Barski, Cuddapah et al. 2007). As investigations into the chromatin state and binding proteins associated with CpG islands are limiting (Tazi and Bird 1990), the use of such new technologies will not allow such a study to be carried out on a genome wide level. The most challenging aspect of the determination of CGI binding proteins lies not in the detection of the proteins themselves but in the purification of pure and stable CpG island chromatin in the first place. As CGIs are rich in non methylated CpG dinucleotides, purification strategies were devised to take advantage of this trait. This was accomplished through the use of methyl sensitive restriction of chromatin using enzymes which contain CpGs in their recognition sequences as described by Tazi and Bird (Tazi and Bird 1990).

3.2 Digestion of chromatin with methyl sensitive restriction enzymes

In order to generate a pure population of CGI chromatin these regions of interest must first be separated away from the bulk of the genomic material. An approach was devised which was heavily reliant on the earlier work of Tazi and colleagues (Tazi and Bird 1990). This method took advantage of the methyl specific DNA digestion imposed by certain restriction enzymes which recognised and cut only at CpG rich sequences, resulting in the liberation of CGI chromatin from the bulk of the genome. Methyl sensitive restriction enzymes such as HinP1 (recognition sequence GCGC) or HpaII (CCGG) can only cut the DNA when the cytosine in the CpG dinucleotide is in an unmethylated state, resulting in the digestion of CpG islands (Figure 3.3-1, A). As digests are carried out in the context of chromatin many of the enzymes target sites will be occluded by the presence of nucleosomes. However, sites between nucleosomes or in nucleosomal free regions over trascriptional start sites will be readily digested by these enzymes. The smallest DNA fragments generated by this digestion range in size from a few base pairs to around 60 nucleotides (nts) and correspond to the non-nucleosomal regions of unmethylated DNA. The next in size are the mono-, di- and trimeric nucleosomal fragments where digestion has taken place around the nucleosome in "unprotected" DNA sequence. Any of the bulk genomic material, which is mainly methylated, will be significantly larger in size allowing for the separation from the small CGI fragments. A set of restriction digestions using the non-methyl sensitive CG recognising enzymes MseI (TTAA) and AluI (AGCT) provide an ideal control experiment as they cut throughout the genome without a preference for CpG island sequences.

As CGIs only contribute to a small proportion of the genome (1-2%), it will be difficult to visualise the fragments released upon digestion by standard ethidium bromide staining. To overcome this problem the DNA fragments were end labelled post digestion with radioactive α^{32}P dCTP nucleotides using Klenow fragment of the DNA polymerase I enzyme. Incorporation of the radioactive base into these fragments allowed for detection by radiation sensitive phosphor screens after gel electrophoresis. Early work using HeLa cells as a substrate was not successful at visualising liberated CGI fragments due to the fact that a fraction of cells in culture routinely undergo apoptosis. As these

apoptotic cells contain degraded DNA this severely raises the background noise seen in the radioactive signals. A new approach using nuclei from mouse brains as a starting point for CGI liberation gave far clearer results (similar to that in Figure 3.4-1).

3.3 Purification of CGI chromatin through sucrose gradient ultracentifugation

As it was now possible to digest the CGI compartment of the genome away from the bulk through methyl sensitive restriction, the next step was to isolate and purify this fraction. The fact that the CGI fragments were much smaller (ranging from <60bp to mono- di- and tri nucleosomal in length) than the remainder of the genome meant that these CGI fractions could be selected for based on their relative sizes. One way in which to accomplish this is by the centrifugation of digested material on a sucrose gradient (Figure 3.3-1). Chromatin is added to columns in which the sucrose concentration is set in a gradient and ultracentrifugation applied which results in the separation of the smaller CGI fragments away from the larger bulk chromatin. Initial experiments were carried out on gradients ranging from 15% to 30% sucrose with later gradients spanning 5% to 30% sucrose. Digested chromatin was gently layered on top of sucrose gradient columns before ultracentrifugation at 20,200g to resolve fragment densities. Once spun, fractions were collected and either analysed for DNA (phenol-ETOH extractions) or protein content (TCA precipitation).

3.3.1 Optimisation of the sucrose gradient for purification of CGI chromatin

Primary analysis was carried out by investigating the distribution of DNA across the sucrose gradient fractions. Purified DNA fragments from the gradient were visualised on a gel through radioactive end labelling α^{32}P dCTP using Klenow fragment of DNA polymerase I as before. The first set of DNA fragments were taken from a 15-30% sucrose gradient and revealed that the small CGI fragments can indeed be separated away from the rest of the genome (Figure 3.3-1). The results of the methyl sensitive (CGI liberating) HinP1 digestions reveal the presence of the small non nucleosomal fragments in fractions 1 and 2 at the top of the centrifugation tube (Figure 3.3-1, B). Fractions 2-5 contain the mono- nucleosomal fragments, with di- and trinucleosomal fragments seen in fractions 3-5 and 4-5 respectively. In fractions 6-9 faint bands are seen corresponding to mono-, di- and

trinucleosomal fractions. The higher fraction numbers correspond to collections from nearer the bottom of the gradient and contain heavy fragments of chromatin. As expected these fractions lack nucleosomes but associate with high molecular weight DNA. At the time of analysis the pellet at the bottom of the gradient containing the majority of very large material was not analysed.

As a control experiment, nuclei digested with Mse1 were run on a 15-30% sucrose gradient. Subsequent precipitation and radioactive end labelling of the DNA reveals that a distinct pattern of fragments are generated in comparison to HinP1 digested material (Figure 3.3-1, B). In contrast to the HinP1 digests, the first 9 lanes of the MseI sucrose gradients contained far lower amounts of nucleosomal material. The majority of the material in this sucrose gradient was found from fractions 10 onwards due to the larger sized of fragment generated through the MseI digestion.

These 15-30% sucrose gradients revealed that some of the digested material was being found in fraction 1 and 2 at the top of the gradient (fractions 1&2; Figure 3.3-1, B). As the uppermost fractions tend to contain large amounts of nuclear proteins this presents a problem of protein contamination. The sucrose gradients were run once more this time over a 5-30% range of sucrose concentrations with the aim of shifting the distribution of these CGI fragments away from the top of the gradient. Ultracentrifugation of these gradients resulted in a similar pattern of fragments visualised to the earlier 15-30% gradients however the CGI fractions were now seen to span fractions 3-9 (Figure 3.3-1, C). On these gradients the non-nucleosomal fragments were now seen to be enriched in fractions 3 and 4 whilst the mono-, di and trinucleosomal fragments are seen primarily enriched in fractions 5-9. In this set of experiment the pellet found at the bottom of the gradient was also included and as expected contained the majority of the DNA corresponding to the bulk of the genome ("pellet", Figure 3.3-1, C). The relative abundance of the DNA in the pellet is unsurprising when one considers the fact that CGIs represent around 1% of the genome. As a control the MseI digest was run over a 5-30% sucrose gradient and once again showed no preference for smaller fragments, with the majority of fragments enriched in the mid range of fractions and in the pellet (data not shown).

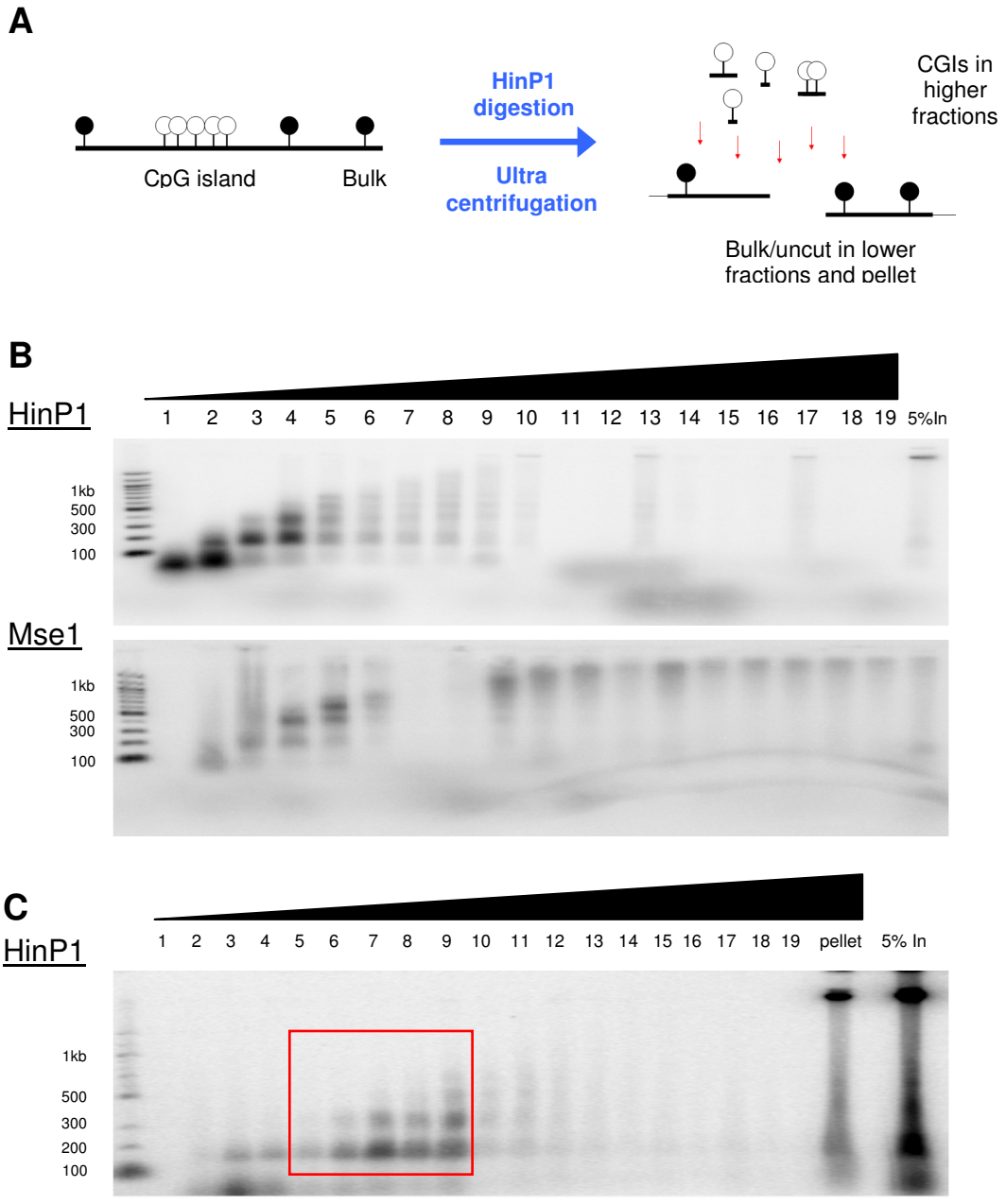

Figure 3.3-1. Methyl sensitive restriction digestion and Sucrose gradient purification of CGI chromatin. A. Non methylated CGI chromatin can be digested into nucleosomal fragments through the methyl specific restriction enzyme, HinP1. These are separated from non CGI chromatin based on the sedimentation of resulting fragments after ultracentrifugation. Mse1 (not shown) is used as a control for non CGI digests. **B.** DNA profiles across the HinP1 and Mse1 15-30% sucrose gradients. Typical CGI sizes correspond to non nucleosomal (< 100 bp), mono-,di- and tri- nucleosomal (~150bp, ~300bp and ~450bp). Fraction number is indicated along the top of the gel from the highest fraction (1) to the lowest (19 and pellet) along with 5% input DNA (5% in). Sucrose gradient density represented by the black slope. Several of the lanes appear to lack any signal, however this is likely due to a loss of DNA during precipitation. **C.** HinP1 5-30% sucrose gradients give a greater separation of CGI fragments away from the uppermost fractions (red box) whilst the majority of DNA is seen in the pellet.

3.3.2 Verification of sucrose gradient purification procedure

In order to verify that the predicted CGI fractions do indeed contain CpG island sequence, PCR reactions were carried out across the 15-30% gradients for two known unmethlyated CGIs *Oprd1* and *Ccne1*. HinP1 digests reveal amplification of both CGIs in lanes 7-10 (Figure 3.3-2, A), corresponding to the predicted CGI fractions in 15-30% gradients (fractions 5-9; Figure 3.3-1, B upper panel). As the PCR product is around 200bp in length it will not amplify from fractions containing only the non nucleosomal (~60bp in size) or mono nucleosomal (~146bp) fragments.

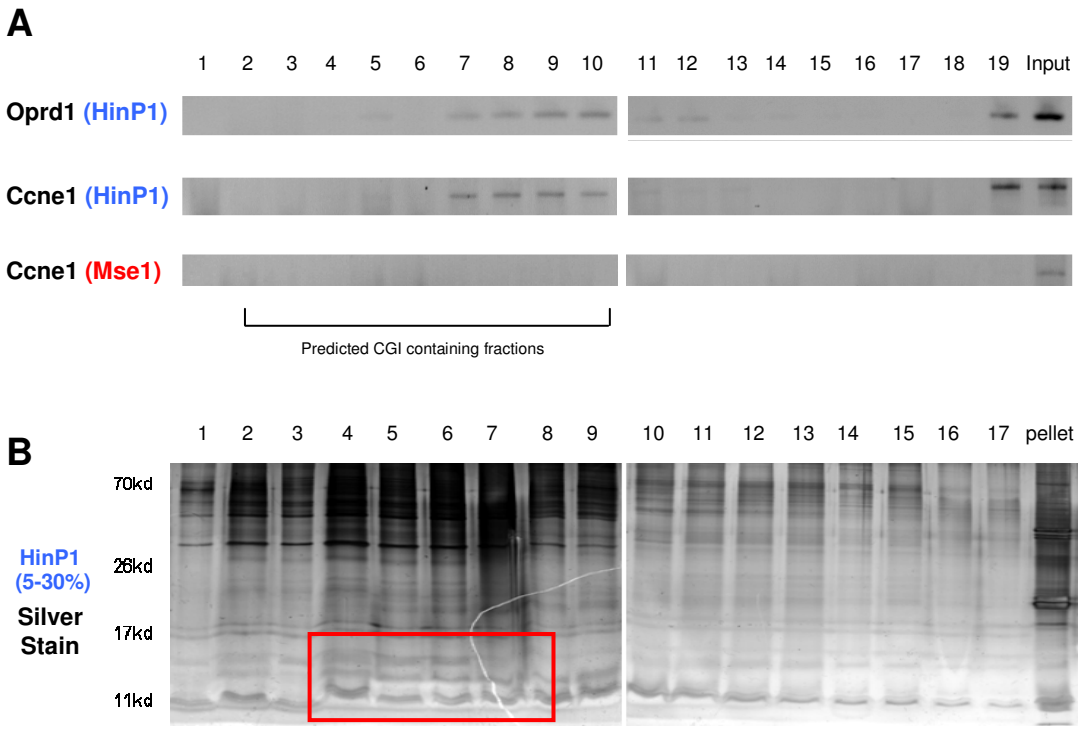

Figure 3.3-2. Verification of CGI liberation through sucrose gradient ultracentrifugation. A. PCR for the distribution of non-methylated CGI across the 15-30% sucrose gradient. Amplification of the Oprd1 and Ccne1 CGIs over the HinP1 gradient was compared to the amplification over an Mse1 gradient. The HinP1 gradient shows the presence of CGIs over the same fractions predicted to contain Islands (fractions 2-10; figure 3.3-1 B), with a minority still in the largely uncut fraction 19 (presumably due to inefficient digestion). In contrast the Mse1 digested sucrose gradient shows a lack of CGI enrichment over the fractions. **B.** whole protein extracts were made across a 5-30% HinP1 sucrose gradient to test for protein content and distribution. Fractions corresponding to those enriched for CGI nucleosomes are seen to contain higher levels of histones by silver staining (red box). However as many of the lanes across the gradient appear to contain similar bands this may indicate background protein contamination.

By contrast analysis across Mse1 digested gradients reveals that no *Ccne1* PCR product is found in the fractions, with the exception of the pellet which will contain all non digested DNA including CGIs (Figure 3.3-2, A).

Although CGI DNA appeared to be successfully liberated from the bulk through sucrose gradient centrifugation, the distribution of proteins across the fractions were investigated to ensure that stable chromatin can be purified and not just DNA. Equal volumes of TCA precipitated material were resolved on a SDS polyacrylamide gel prior to silver staining (see materials and methods section 2.2.5). Although high levels of proteins were seen in each of the lanes making it hard to define any unique CGI proteins, levels of histones were indeed enriched here as expected from the earlier DNA studies (Figure 3.3-2, B). The finding that each lane contained so many proteins was somewhat surprising as it was hoped that this approach would result in the visualisation of bands representing CGI specific binding factors. In fact this result highlights the potential problems with the sucrose gradient approach to liberate CGI chromatin which can be explained further after mass spectrometry.

The fractions corresponding to the CGI fragments were analysed by mass spectrometry. Fractions 5-9 of the HinP1 (suspected CGI chromatin) and MseI (control chromatin) digested nuclei (Figure 3.3-1, C) were run a small distance on a SDS polyacrylamide gel. These gel lanes were then cut out, trypsinised and analysed by mass spectrometry. Resulting lists of proteins attained for HinP1 (CGI chromatin), Mse1 (bulk chromatin) and uncut chromatin (as a control) failed to identify many CGI specific proteins. HinP1 fractions identified 841 proteins whilst the uncut control found nearly identical numbers of proteins. Comparative analysis of the three data sets revealed that 87 proteins were unique to the HinP1 digested chromatin fragments (Figure 3.3-3). However, closer inspection of these proteins found that around sixty of these appeared to have no relevance to CGIs or promoter function whilst seven were entirely unknown in function. Of the remaining twenty proteins twelve were transcription factors (e.g. LBX1) whilst three were associated more directly to transcription (e.g. Polymerase II polypeptide C). Other proteins of interest include the DNA replication origin licensing factor MCM6 as well as three RNA binding proteins.

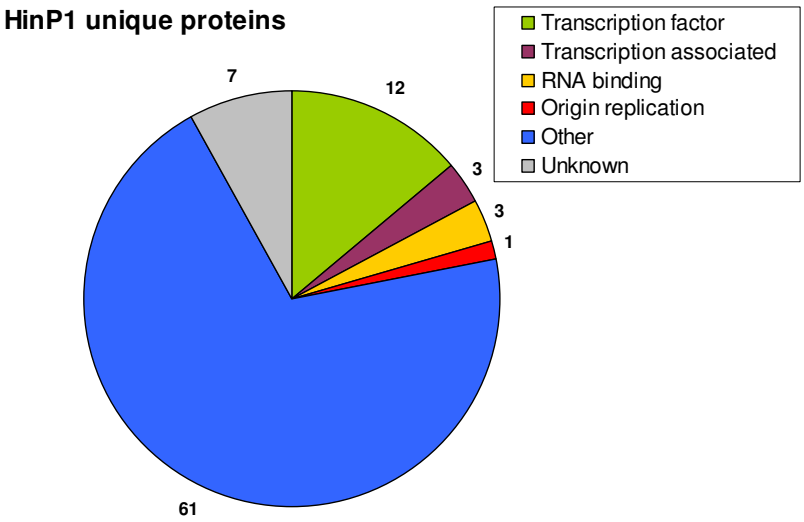

Figure 3.3-3. HinP1 specific proteins as revealed through mass spectrometry. The 87 proteins found to be unique for HinP1 digests (CGI specific) are represented on a pie chart, grouped by their cellular function. The majority (61 and 7) were either unidentifiable proteins or those with non-CGI or non-promoter/transcriptional links.

Due to the high number of non-CGI related proteins present in both the HinP1, Mse1 and uncut samples, this approach was unable to accurately determine the protein compliment at CGIs. The problem of protein contamination could be explained if the chromatin is unstable in the sucrose environment and the proteins dissociate from the DNA, sedimenting to their respective densities. This would explain why proteins such as MeCP2 (a protein shown to bind to methylated CpG rich DNA) is seen in all three samples (data not shown). It is worth noting that this experiment was carried out prior to the identification of the somewhat more efficient and simple restriction enzyme digestion method for the purification of CGI chromatin (chapter 3.5). Repeating this analysis with this improved purification method may reveal novel CGI binding factors without the background contamination.

3.4 Salt wash and methyl sensitive restriction to liberate CGI chromatin

The results of the mass spectrometry reveal a high level of background protein contamination. To ensure that any non-specific background proteins were removed an approach was taken to liberate and purify stable CGI chromatin. This was accomplished through the combination of a high salt wash protocol (Dignam, Lebovitz et al. 1983) with methyl sensitive restriction. Mouse brain nuclei were washed thoroughly in moderate levels of salt (0.3M NaCl) to remove non-specific and low affinity chromatin binding proteins before incubation with the restriction enzyme of choice (be it the methyl sensitive HinP1 or insensitive MseI). During this incubation period, the restriction enzymes are able to gain access to and selectively digest the chromatin in the nucleus. The resulting small CGI chromatin fragments are released into the supernatant and can be separated from bulk and uncut chromatin by centrifugation at low speeds. A set of calibrations determined the optimal speed for the centrifugation to be around 3,800g in order to best separate the smaller CGI fragments (non-, mono-, di- and trinucleosomal fragments) from the bulk (data not shown). Precipitated DNA from the supernatants of the digested nuclei were end labelled as before through the incorporation of a radioactive α^{32}P dCTP (see section 3.2 and 3.2). Half of this labelled DNA was then digested further by another methyl sensitive restriction enzyme HpaII (cleavage site CCGG). As this DNA is now deproteinised (as opposed to the earlier HinP1 digests which were carried out in the context of chromatin), novel non methylated CpG sites are now able to be digested by the HpaII enzyme (Figure 3.4-1, A). If CGIs are present this should reduce the size of the previously digested DNA fragments further, appearing as a "collapse" of the nucleosomal pattern as shown previously by Antequera and collegues (Antequera, Macleod et al. 1989).

The results show that the HinP1 digested material is indeed rich in CGI sequences. Firstly HinP1 digestion alone gives the characteristic non nucleosomal, mono-, di- and trinucleosomal pattern found associated with successful CGI release. In contrast the MseI digested nuclei do not give this pattern, instead generating a range of band sizes representing general digestion throughout the genome (Figure 3.4-1, A). Furthermore, HinP1 (CGI) chromatin is collapsed further upon HpaII digestion revealing that this DNA is indeed unmethylated and CpG rich. By contrast the MseI digests is unaffected by subsequent HpaII digestion, revealing that this material is devoid of unmethlyated CpG rich DNA.

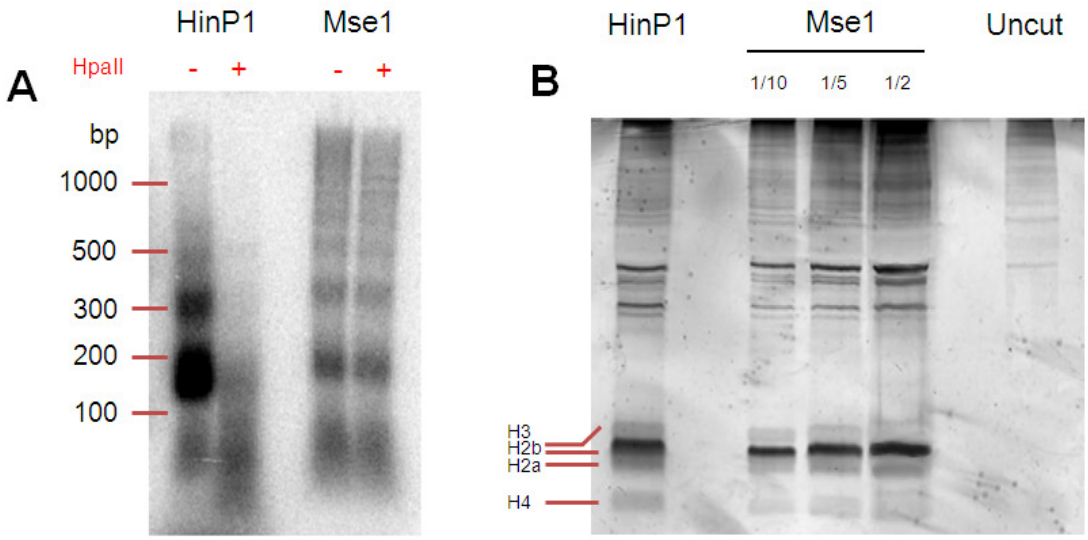

Figure 3.4-1. DIGNAM salt wash and methyl sensitive digestion of nuclei successfully liberates CGI chromatin. A. Fragments of DNA from the digested nuclei are radioactively end labelled to allow visualisation on an agarose gel. HinP1 digests reveal the expected pattern of nucleosomal digestion (HinP1, HpaII –ve lane) implying the successful liberation of CGI material. As a control, Mse1 digestions give rise to genomic digestion patterns (Mse1, HpaII –ve lane). Deprotinization post liberation followed by further digestion with the methyl sensitive restriction enzyme HpaII reveals that the suspected CGI fraction (HinP1 digests) can be digested further due to exposure of further non methylated CpG sequences after protein loss. This collapse is not seen in the Mse1 digests implying that a pure population of CGI chromatin is indeed liberated through HinP1 digestion of nuclei. Marks corresponding to a 100 base pair ladder are shown on the left. **B.** Total protein content from HinP1, Mse1 (at a range of concentrations) and mock cut reactions analysed by silver stained polyacrylamide gels. Digests reveal a strong proportion of the proteins liberated are histones (predicted histone sizes are shown on the left). Additionally it is apparent that the Mse1 digested nuclei liberate around five times as much protein as the HinP1 digests.

However, as seen with the sucrose gradient experiments, it is important to test the protein compliment of this purified material to ensure that stable chromatin is being present. Silver stained gels containing proteins present in the supernatants of HinP1, Mse1 and uncut control nuclei reveal far lower overall protein levels when compared to those seen in the sucrose gradients (Figure 3.4-1, B). The Dignam salt washes have resulted in a loss of protein contamination as seen in the uncut nuclei lane of Figure 3.4-1, B. It appears from these results that the MseI digestion releases around five times as much protein into the supernatant, mainly due to the fact that this enzyme can cut more frequently than HinP1.

3.5 Summary of CGI chromatin isolation technique

This adaptation of the original method of isolating CGI chromatin as described by Tazi and collegues (Tazi and Bird 1990) produced far clearer results. The work by Tazi and Bird showed similar patterns of DNA digestion and collapse of CGI rich DNA to that seen in figure 3.4-1, A. However as the work presented here was also interested in CGI specific proteins this method had to be adapted to remove the original sucrose gradient size selection (which was found to be determental for protein stability within the context of chromatin) for an alternative digestion and salt wash method (Dignam, Lebovitz et al. 1983).

This adapted method of isolating CGI chromatin has greatly reduced the level of "background" protein contamination found when samples were subjected to sucrose gradient centrifugation (figure 3.3-2,B), although it is hard to see any differences between the proteins present in the "CGI" and "bulk" chromatin fractions by looking at a silver stained gels alone (figure 3.4-1,B). Further investigations into the specific histone modifications enriched within the CGI fraction through techniques such as immunoblotting should help to better understand the chromatin modifications associated with CGIs. Potential CGI specific binding factors can also be tested for their distributuion between the HinP1 and MseI digested material. Overall this adaptation on a classical method has successfully liberated CGI chromatin from the rest of the genome.

Chapter 4: Investigating the Protein Compliment of CGIs

Historically the majority of work on CpG islands has focused primarily on the study of the DNA. The hypothesis that these non-methylated CpG rich stretches of DNA are associated with a unique set of proteins has not been thoroughly tested. However, the identification of any CGI specific proteins could go some way to explaining any functions associated with these loci. Additionally any proteins found to bind over CGIs may be responsibly for the maintenance of these unmethylated sites within a largely methylated genome.

Progress in such investigations have been hindered by the lack of powerful enough analytical techniques. However, recent advances in the sensitivity of mass spectroscopic analysis and the advent of high resolution whole genome sequencing may allow the identification of CpG island binding proteins. Whole genome sequencing technology has already led to a plethora of studies investigating the distribution of histone modifications and DNA binding proteins on a genome wide scale. This technology has allowed the rapid generation of high resolution binding patterns for multitude of proteins in a wide variety of organisms (Barski, Cuddapah et al. 2007; Cronn, Liston et al. 2008; Whiteford, Skelly et al. 2009). As the earlier mass spectrometry approaches at determining CGI binding proteins had failed a more candidate driven approach was devised to identifying a true set of CpG island binding factors.

4.1 Comparisons of CGI and genomic chromatin

4.1.1 Candidate driven analysis of Global CGI protein compliment

As CGI chromatin had been successfully liberated from the bulk through a series of salt washes followed by methyl sensitive restriction enzyme digestion (section 3.5), this chromatin could be used to analyse the distribution of candidate CGI binding factors and histone modifications though western blotting.

In order to correctly test for enrichment of any factors, each of the protein samples had to be calibrated to ensure equal loading. Histone H3 was used due to its abundance, ease of detection and relative consistency across the genome. However, nucleosomal depletion is often seen over the TSS of many genes in the mammalian genome (Heintzman, Stuart et al. 2007; Ozsolak, Song et al. 2007; Ramirez-Carrozzi, Braas et al. 2009). As CGIs co-localise to around 60% of annotated promoters this may result in lowered levels of H3 within population of CGI chromatin. However the use H3 as a loading control allows the direct comparison of histone tail modifications between CGI and bulk genomic chromatin.

In order to calibrate the levels of H3 within the CGI and bulk genomic chromatin, a western blot was carried out comparing levels of the protein between HinP1 digested nuclei (CGI fragments) and several dilutions of MseI digested nuclei (control digests; Figure 4.1-1, A). Quantification of H3 signals reveals approximately 5 fold less histone H3 in the CGI chromatin samples compared to the control Mse1 digests. Using this as a control for equal loading, histone modifications can therefore be confidently compared between CGI and bulk chromatin. The relative abundance of these modifications may lead to a greater understanding of CGI function and maintenance.

Acetylated Histone H3

The first modification tested was acetylated Histone H3, a mark tightly linked to the expression of genes (chapter 1.1.1 section iv). As this modification can be found at lysines 9, 14, 18, 23 and 27 on the H3 tail an antibody was selected which would recognise all such modifications (entitled a "pan-acetyl" H3 antibody). This mark is found enriched over the promoter start sites of active genes (Kurdistani, Tavazoie et al. 2004) it was previously shown to be enriched withing CGI chromatin (Tazi and Bird 1990) . Accoridingly the modification was seen enriched in the CGI chromatin relative to bulk chromatin (Figure 4.1-1, B). Acetylation of the H3 tail is not only confined to CGI chromatin but is found in the bulk fraction, possibly due to the transcriptionally active non-CGI genes (around 35% of all annotated mouse genes).

H3K4 methylation

The next histone modification tested was that of H3K4 methylation. High resolution studies have shown that mono methylated H3K4 modification peaks between -900 and +1000bp over promoters whilst, di- methyl is found between -500 and +700bp at the promoters of active genes (Figure 1.1-2) (Barski, Cuddapah et al. 2007). Interestingly the tri- methylated version of H3K4 is seen to peak strongly directly over the transcriptional start site of a gene and this occurs regardless of the underlying transcriptional state although typically with lower values at inactive or lowly transcribed genes (Barski, Cuddapah et al. 2007). As CGIs are often found associated with the promoter regions of genes these H3K4 methyl modifications are predicted to be found in the HinP1 digested chromatin. Indeed western blots for both H3K4me2 and me3 reveal that both are enriched in the CGI chromatin relative to the bulk (Figure 4.1-1, B)

Other histone modifications

The next set of histone modifications investigated were those typically found associated either within the bodies of genes or with inactive regions of chromatin. The histone mark H3K36me3 for example, is a modification associated with transcriptional elongation. This mark is found to be depleted over promoter regions but enriched over the gene body, mainly in highly transcribed genes (Barski, Cuddapah et al. 2007). It is believed that this modification works in conjunction with acetylation to allow the RNAP II complex to pass along the gene body during transcriptional elongation (Lee and Shilatifard 2007). As this modification is known to be absent from promoter sites it is not surprising to find that it is depleted in the CGI chromatin relative to bulk (Figure 4.1-1, B). Interestingly this mark is not entirely absent from purified CGI chromatin, possibly due to the fact that around 25% of all CGIs in the mouse are intragenic and thus found within gene bodies along with the H3K36me3 modification.

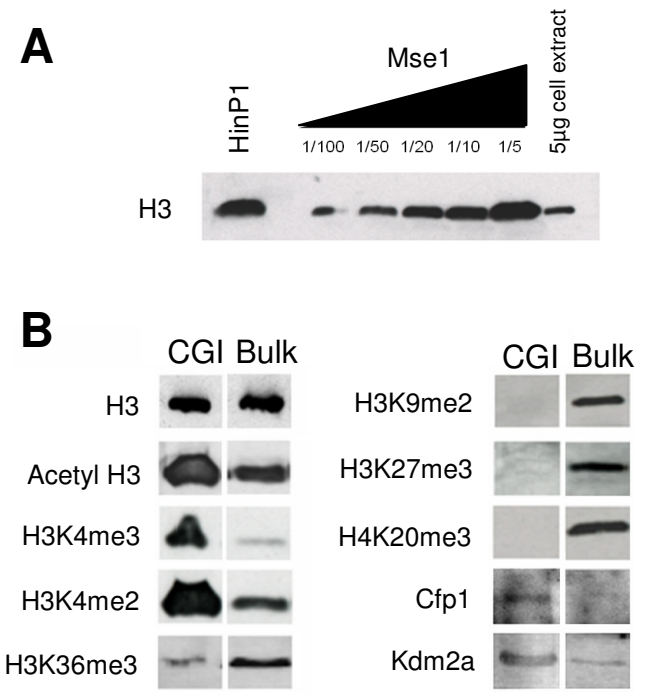

Figure 4.1-1. The distribution of histone modifications between CGI and bulk chromatin.

A. Western blot for histone H3 levels to calibrate the equal loading of protein samples. Mse1 digests liberate around 5 times as much material as HinP1 digests. Samples were calibrated as relative volumes of TCA precipitated material (from equal numbers of nuclei). Dilution values are displayed across the top of the western blot. As a control 5µg of whole cell extract was loaded in the far right lane. **B.** Selected histone modifications were compared between liberated CGIs (HinP1) and bulk chromatin (Mse1). Equal loading was established by H3 levels (left group, top panel). The "active" histone modifications including acetylated H3 and H3K4me2 & 3 were all significantly enriched within the CGI fraction relative to the bulk. Non-active marks such as H3K36me3, H3K9me2, H3K27me3 and H4K20me3 are all depleted or completely absent from CGIs. Two candidate CGI binding factors, Cfp1 & Kdm2a, reveal enrichment within CGI chromatin

Probing for the distribution of several other histone modifications which tend to have links to transcriptionally inactive chromatin gave similar results. The modifications H3K9me2, H3K27me3 and H4K20me3 have all been shown to be enriched at the promoters of inactive genes and are associated with the formation of the heterochromatic regions of the genome. Western blot analysis using antibodies to these modifications reveal a striking depletion of all three of these marks within CGI chromatin when compared to the bulk of the genome (Figure 4.1-1, B).

In summary the CGI chromatin is enriched in the histone tail modifications typically associated with gene activation such as H3 acetylation and H3K4me2 and me3. Furthermore this CGI chromatin appears to be depleted in the transcriptional elongation mark of H3K36me3 as well as the silencing modifications of H3K9me3, H3K27me3 and H4K20me3. Therefore CGI chromatin does indeed differ from the rest of the genomic material with regards to the modifications of the associated histones.

CXXC proteins at CpG islands

After the histone modifications associated with CGIs had been investigated the next step was to investigate this chromatin for specific binding proteins which may play a role in the initiation or maintenance of the CpG islands state. One particular property of the CGI that may be important in the binding of proteins is that the DNA sequence is extremely CpG rich. The MBD group of proteins are known to bind to methylated CpG dinucleotides through interactions between the cytosine base and a conserved methyl binding domain. It would therefore be interesting to search for a group of proteins containing any domain with specific affinity for non-modified CpG dinucleotides. One such candidate is the CXXC domain which is found in a group of around ten proteins, many which have roles in the regulation of chromatin state (see chapter 1.4). This zinc finger domain has been shown to bind to unmethylated CpG rich DNA oligonucleotides through interactions with zinc cations (Voo, Carlone et al. 2000; Lee, Voo et al. 2001). Furthermore this domain was used in the Bird laboratory as a means of generating CGI data sets by CAP technology (chapter 1.3.2). These findings indicate that proteins containing this CXXC domain may have important roles at CGIs.

The first of the CXXC proteins investigated was a protein called CpG binding protein, Cfp1 (also known as CXXC1 or CGBP) which had previously been shown to be a component of the hSet1 H3K4 HMTase complex (Lee and Skalnik 2005; Lee, Tate et al. 2007). Western blot analysis reveals that Cfp1 is enriched within CGI chromatin (Figure 4.1-1, B). A second protein containing the CXXC domain, Lysine specific demethylase 2A, KDM2a (also known as Jhd1a) was also tested for its global distribution. Kdm2a is the enzyme responsible for the removal of methylation at lysine 36 on the histone H3 tail. As discussed above, this mark is involved in transcriptional elongation and is depleted over gene promoter regions. Like Cfp1 this protein is greatly enriched within the CGI fraction of the genome (Figure 4.1-1, B). However low levels of Kdm2a are found outside of CGI regions, possibly

as non-CGI promoters (which presumably will be H3K36me3 free and KDM2A positive) will be found in bulk genomic fractions.

4.1.2 Database analysis of histone modifications associated with CGIs

In order to verify the results of the western blots, the findings were compared to high throughput histone modification ChIP data sets. Using the online database of epigenetic marks collected by the ENCODE consortium (Encyclopedia Of DNA Elements: http://genome.ucsc.edu/ENCODE), CpG islands were investigated for the same histone marks previously probed experimentally. The ENCODE database displays the profiles of histone modifications across a small region of the genome which will invariably contain CGI chromatin. 60 randomly selected promoter CpG islands were investigated for associated histone modifications. Human lymphoblast cells were selected as this cell type had the most thorough analysis of histone modifications documented within the ENCODE database. Histone modifications at these 60 CGIs were classed as either “present” (where peaks of modifications were strongly associated to CGIs), “low” (where trace or unconvincing peaks of modifications existed) or “absent” (where CGIs were totally devoid of the mark of interest). Although limited, this analysis allows for comparisons to be drawn between the western blot results and a high throughput data set.

	H3ac	H3K4me2	H3K4me3	H3K36me3	H3K9me3	H3K27me3
Present	60	88.4	91.7	6	0	0
Low	15	11.6	8.3	14	8.3	20
Absent	25	0	0	80	91.7	80
WB (CGI)	Present	Present	Present	Low	Absent	Absent

Table 4.1-1 –Distribution of histone modifications at predicated CGIs through database mining. The table represents the percentages several histone modification across the 60 CpG island promoters analysed. "Present" refers to modifications which peak at moderate to strong levels over CGI regions whilst "low" refers to small and possibly non-significant peaks. "Absent" marks were completely lacking across the CGIs investigated. Highlighted boxes denote the higher percentage group, thus a consensus for the mark of interest, e.g. H3K4me3 is typically found present at moderate to high levels in promoter CGIs whilst H3K4me1 is generally found only at low levels. As a comparison the results of CGI chromatin from the western blot in figure 4.1-2 are shown below ("WB CGI").

Overall this analysis verifies the results of the global western blot experiments (figure 4.1-1) (Table 4.1-1). The methyl H3K4 modifications were seen to be tightly associated with CGIs increasing in association depending on the level of lysine methylation present (mono < di- < trimethylation). None of the 60 promoters CGIs investigated lacked the di- or tri- methyl mark. Acetylated H3 was seen to be somewhat enriched at CGIs; however a quarter of islands lacked this modification (Table 4.1-1). These findings raise the question as to what is maintaining this H3K4 mark at the CpG islands at promoter loci (which vary in their levels of H3 acetylation). Analysis of the histone modifications which are generally thought to be associated with an inactive chromatin state also follow the findings of the earlier western blot study. 80% of the CGIs investigated were completely devoid of the H3K36me3 modification previously show to be absent at CGIs. Other silencing marks such as H3K9me3 and K27me3 also follow this pattern with 92% and 80% of CGIs completely lacking these modifications. It is worth noting that all of these three modifications were seen at low levels at around 10-20% of CGIs however this may be due to the somewhat loose nature of categorisation used to define the classes.

4.2 Profiling of promoter CGIs for associated proteins

As several histone modifications and CXXC proteins were shown to be enriched at CpG islands on a global level, individual islands were subsequently inspected in greater detail through the use of chromatin immunoprecipitation followed by quantitative PCR (ChIP PCR). This approach allows the quantification and distribution of candidate proteins and histone modifications across regions of interest such as a CGI at a particular locus.

The recent generation of CGI data sets in the Bird laboratory reveals that around 50% of identified CGIs are found associated with the 5' end of an annotated gene with the remaining 50% at non promoter defined regions (Illingworth, Gruenewald-Schneider et al. 2010). Approximately half of these are intra-genic and half inter-genic. Representatives of all three classes were analysed by ChIP-PCR and the results compared. Four promoter CGIs were investigated, including "typical" non methylated promoters as well as a methylated CGI. The profiles shown are results from a single experiment which were "typical" of the profiles seen over at least three ChIP-PCR experiments. There were often variations in the immunoprecipitation values between experiments leading to large error bars when the data was combined. Although these IP values differed the patterns seen did not vary and the overall enrichment values of peaks relative to flanking regions remained consistent.

Seven promoter CGI were tested in mouse brain. Six of these were non-methylated in mouse brain tissue: genes for Brain-derived neurotrophic factor (*Bdnf*) which contains 2 CGIs, β-Actin (*Actb*), *C-Myc*, Opiod receptor delta 1 (*Oprd1*) and Forkhead box D3 (*Foxd3*). In addition to these somewhat "typical" non-methylated CGIs, an island which is constitutively methylated in the brain (*Sycp2*) was also tested.

4.2.1 Cfp1 profiles over promoter CGIs in the mouse brain

The CXXC protein of Cfp1 was the first protein investigated and was seen to bind over four of the six non methylated CGIs tested. Little or no binding was seen over the methylated CGI at the promoter of

the *Sycp2* gene and the CGIs at the *Oprd1* and *Foxd3* promoters (Figure 4.2-1). Typical profiles for this protein reveal low levels of background signal over the flanking regions with discrete peaks seen peaking over the regions corresponding to predicted CGIs. Focussing on the peaks of Cfp1, typical IP efficiencies for Cfp1 peaks were in the range of 0.15% and 0.2% for the *bdnf* CGIs, 0.5% for *actb* and 0.3% for c-*myc*. However, a far more appropriate way to determine the relative enrichment of the protein of interest over CGIs is to compare the binding values at these loci to the surrounding flanking regions. Using this measurement Cfp1 is seen to be enriched 4.2 fold and 3.1 fold over the two CGIs of *bdnf*, 5.2 fold enriched over *actb* and 2.6 fold enriched over *c-myc*. This reinforces the earlier finding that the protein Cfp1 is indeed a CGI binding factor (see section 4.1). As mentioned above, two of the 6 non methylated CGIs (*Oprd1* and *Foxd3*) appear to lack defined peaks of Cfp1. This raises the possibility that Cfp1 is not an ubiquitous CGI binding protein *per se* but can be targeted to these loci under given circumstances.

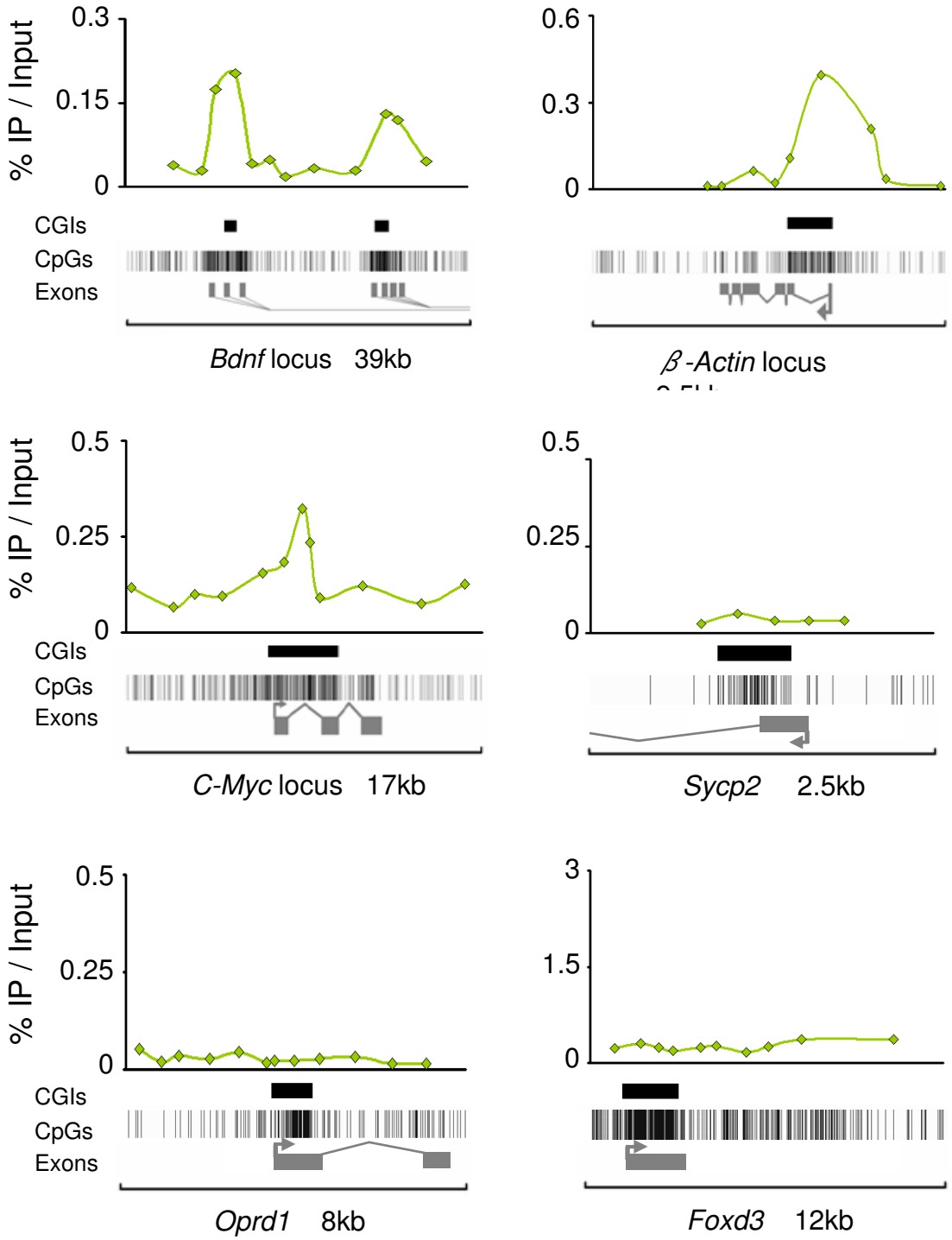

Figure 4.2-1. ChIP-PCR profiles over promoter CGIs for the CXXC protein Cfp1. Each of the 6 panels represents a promoter region. Gene structures are shown below the profiles with grey boxes representing exons ("exons"). CpG density plots ("CpGs") are shown above the gene structure with each vertical line representing a CpG dinucleotide. Refseq derived CGIs are shown above the CpG density plots as indicated by the black boxes ("CGIs"). The enrichment values of the ChIP-PCR results are represented as the % of immunopreciptated DNA enriched relative to the input material ("%IP/Input"). Midpoints of resulting PCR amplicons are plotted along the x axis. From these results, Cfp1 is seen to peak over the islands at *bdnf, β-actin* and *C-Myc* but not at *Sycp2, Oprd1* and *Foxd3.*

The results of Cfp1 Chip over the methylated Sycp2 CGI add weight to the suggested role of Cfp1 as a non-methyl binding protein. This is not entirely unexpected as the CXXC domain within Cfp1 was show previously to only bind to non methylated CpG dinucleotides *in vitro* (Voo, Carlone et al. 2000). However the ChIP results over Sycp2 gives an indication *in vivo* that Cfp1 does not bind to methylated CpG rich islands. A more thorough method of testing the *in vivo* binding behaviour of Cfp1 was carried out later (see section 5.2.2 for more).

4.2.2 H3K4me3 profiles over candidate promoter CGIs in the mouse brain

The active histone modification of H3K4me3 previously found to be globally enriched at CGIs was also seen enriched at several of the candidate islands through ChIP-PCR analysis (Figure 4.2-2). Furthermore H3K4me3 enrichment profiles closely matched those of Cfp1, peaking over the CGIs of *bdnf, actb* and c-*myc*. Enrichment values of binding over the CGIs show strong and well defined H3K4me3 peaks over these regions. For example, the *bdnf* locus shows 13.3 and 6.8 fold enrichment of this modification at the CGIs versus the flanking regions whilst *actb* and *c-myc* reveal 2.9 and 7.25 fold enrichment values at CGIs. The enrichment value of H3K4me3 at *actb* is somewhat reduced due to the fact this modification spans a region greater than the CGI; however this modification is at its highest over the CGI region.

The methylated CGI at the *sycp2* locus shows no discernable peak of the H3K4me3 modification across its length, a finding made all the more interesting through the lack of Cfp1 binding at this loci. Additionally, the other two CGIs which were previously shown to be devoid of Cfp1 were also shown to lack the H3K4me3 mark. As Cfp1 is part of the enzymatic complexes responsible for the establishment of H3K4me3 patterns this provides a possible link between the CXXC protein and the histone modification at CpG islands.

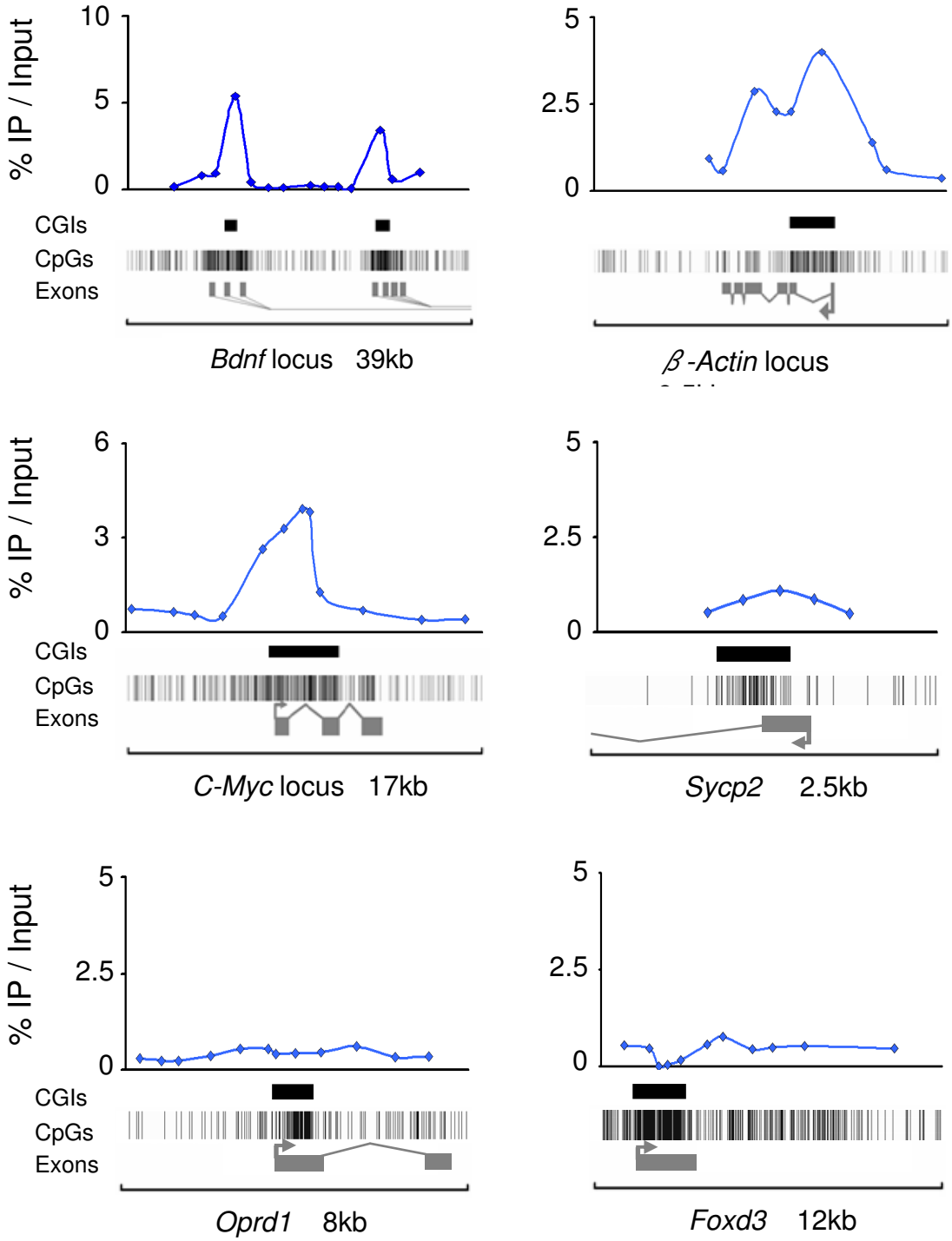

Figure 4.2-2. ChIP-PCR profiles over promoter CGIs for the histone modification H3K4me3.

ChIP-PCR profiling for the active histone modification H3K4me3 reveals a similar pattern of binding over CGIs to that of the Cfp1 protein. H3K4me3 is seen to peak over *bdnf*, *β-actin* and *C-Myc* but not at *Sycp2, Oprd1* and *Foxd3*.

General figure layout and abbreviations follow those outlined in figure 4.2-1.

4.2.3 H3K27me3 profiles over CGIs

The histone modification H3K27me3 is generally associated with "silenced" regions of chromatin set up by the Ezh2 subunit of PRC2 (polycomb repressive complex 2) (Cao, Wang et al. 2002) and is in turn bound by the PRC1 complex (polycomb repressive complex 1) to maintain this silencing. Regions high in this histone modification tend to exist in large domains in contrast to the discrete peaks of the H3K4 methylation states (Boyer, Plath et al. 2006; Roh, Cuddapah et al. 2006; Mikkelsen, Hanna et al. 2008). Genome wide studies have revealed that the majority of this modification is seen to correspond to inactive or "silenced" gene promoters however domains of this modification have also been detected across many intergenic regions of the genome (Zhao, Han et al. 2007).

Typically, non methylated CGIs tend to lack the H3K27me3 modification in differentiated cells with profiles tending to remain low over such loci (Figure 4.2-3). However, profiling of the *Bdnf* locus in mouse brain revealed that both of the CGIs at this promoter contain peaks of H3K4me3 (Figure 4.2-2) as well as H3K27me3. These H3K27me3 peaks are found to have 3 to 3.5 fold enrichment over both of the CGIs relative to the flanking regions. A possible explanation for this is that the brain material used for ChIP contains multiple cells types such as neurons, glia etc and represent a mixed population of cells in which this gene is either on or off (thus H3K4me3 positive/H3K27me3 negative or *vice versa*). However it cannot be ruled out that the two CGIs at *bdnf* are indeed bivalently marked by both H3K4me3 and H3K27me3 modifications.

In contrast the β-*Actin* locus exhibits a far more typical H3K27me3 patterns with around a 4 fold reduction in the modification over the CGI relative to the flanking sites. It must be noted that the signal over this entire locus was nearly ten fold lower than that seen over *Bdnf* further highlighting the low levels of this modification found over this CGI.

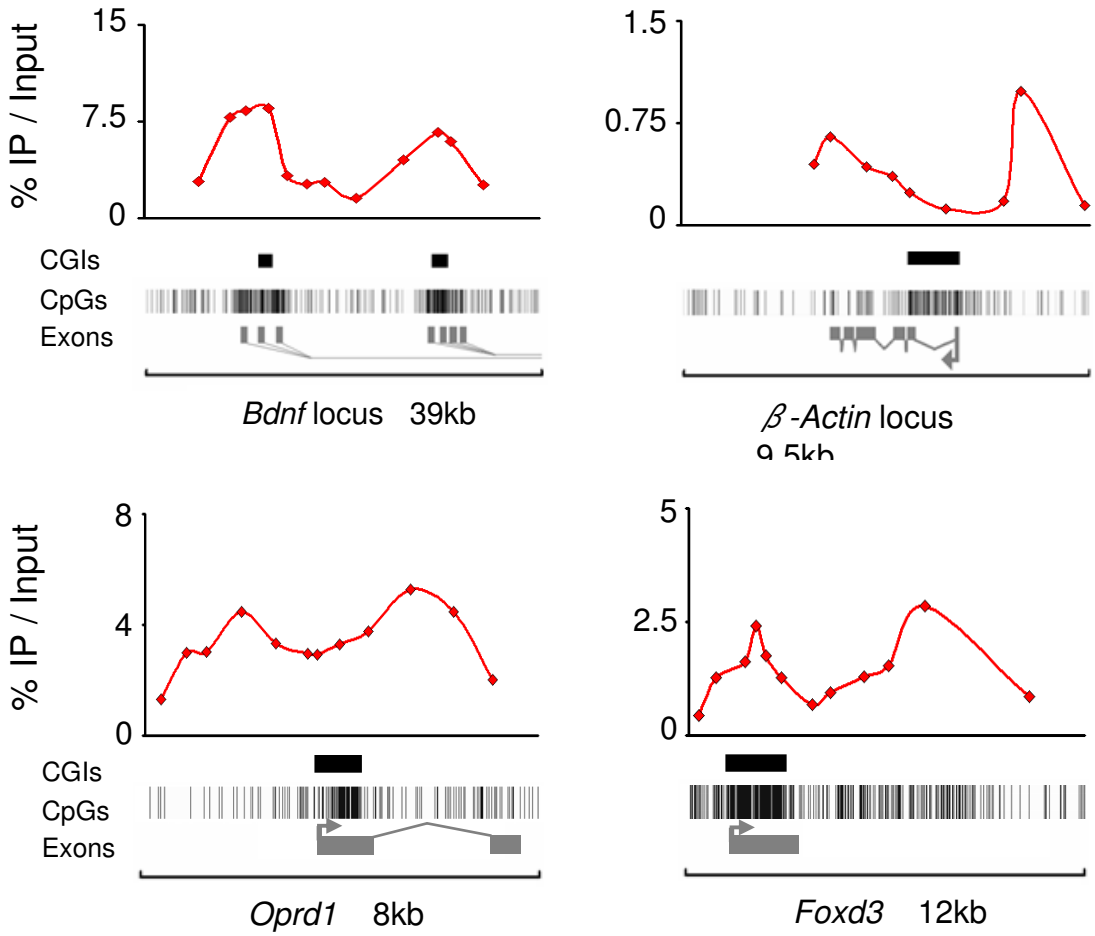

Figure 4.2-3. ChIP-PCR profiles over promoter CGIs for the histone modification H3K27me3.
ChIP-PCR profiling for the silencing histone modification H3K27me3 over 5 CGIs at 4 different promoters. H3K27me3 was seen to peak over *bdnf, Oprd1* and *Foxd3* CGIs but not that at *β-actin.*

General figure layout and abbreviations follow those outlined in figure 4.2-1.

Interestingly the two Cfp1 and H3K4me3 negative islands of *Foxd3* and *Oprd1* reveal high levels of H3K27me3 modified histones. Whether or not this modification directly repels Cfp1 binding and H3K4me3 modification or is a secondary event in the formation of heterochromatin is not yet known. It is possible that the promoters of these two genes were initially bivalently marked with both H3K4me3 as well as H3K27me3 in ES cells, leading to a transcriptionally "poised" state which is lost upon differentiation (Guenther, Levine et al. 2007). Foxd3 is indeed a developmental gene which is not transcribed in the brain and therefore silencing modifications such as H3K27me3 are found here. The links between the H3K27me3 modification and both H3K4me3 and Cfp1 binding are somewhat

complicated by the results at the *bdnf* locus. Although these finding can be explained by assuming a mixed cell population leads to the portrayal of a "bivalent-like" pattern over the CGIs this needs to be tested. Repeating these experiments on a pure population of cultured cells may give rise to a more accurate set of results.

4.2.4 H3K9me3 profiles over promoter CGIs in mouse brain

Both di- and Tri- methylation of H3K9 has been implicated in both heterochromatin formation and gene silencing events (Bannister, Zegerman et al. 2001). Global levels of this modification were shown above to be absent or non detectable within non methylated CGIs (Figure 4.1-2, B). Profiling of this modification over the CGI regions of *Bdnf*, β-*Actin* and *C-Myc* show this modification enriched across all of these loci (Figure 4.2-4). By contrast, the methylated (and silenced) promoter at the *Sycp2* locus exhibited around a 2.5 fold peak of H3K9me3 over the CGI region of this gene with IP efficiencies peaking at around 5% of the input.

The promoter CGIs at *Oprd1* and *Foxd3* which lacked Cfp1 and H3K4me3 but contained H3K27me3 also contain peaks of H3K9me3. Both loci show peaks of this silencing modification although the *Oprd1* locus contains a far less defined peak than the *foxd3* CGI. However, these loci along with the CGI at the *Sycp2* promoter show IP efficiencies which are around ten fold higher than those seen at *bdnf, actb* and *C*-myc. All three of the islands which lack Cfp1 (the methylated *Sycp2* and non methylated *Oprd1* and *Foxd3*) contain peaks of both the silencing histone modifications of H3K9me3 and H3K27me3 whilst the islands that contain this CXXC protein lack these marks and instead containing the active histone modification H3K4me3.

4.2.5 Profiles of other CXXC proteins over CGIs in the mouse brain

Several other proteins contain a CXXC domain homologous to that found in Cfp1 (Figure 1.4-1). Due to the affinity of this domain for non-methylated CpG dinucleotides, these proteins may represent a family of CGI specific binding factors. To determine whether or not these proteins are enriched over promoter CGIs, ChIP-PCR experiments were performed over the regions tested above.

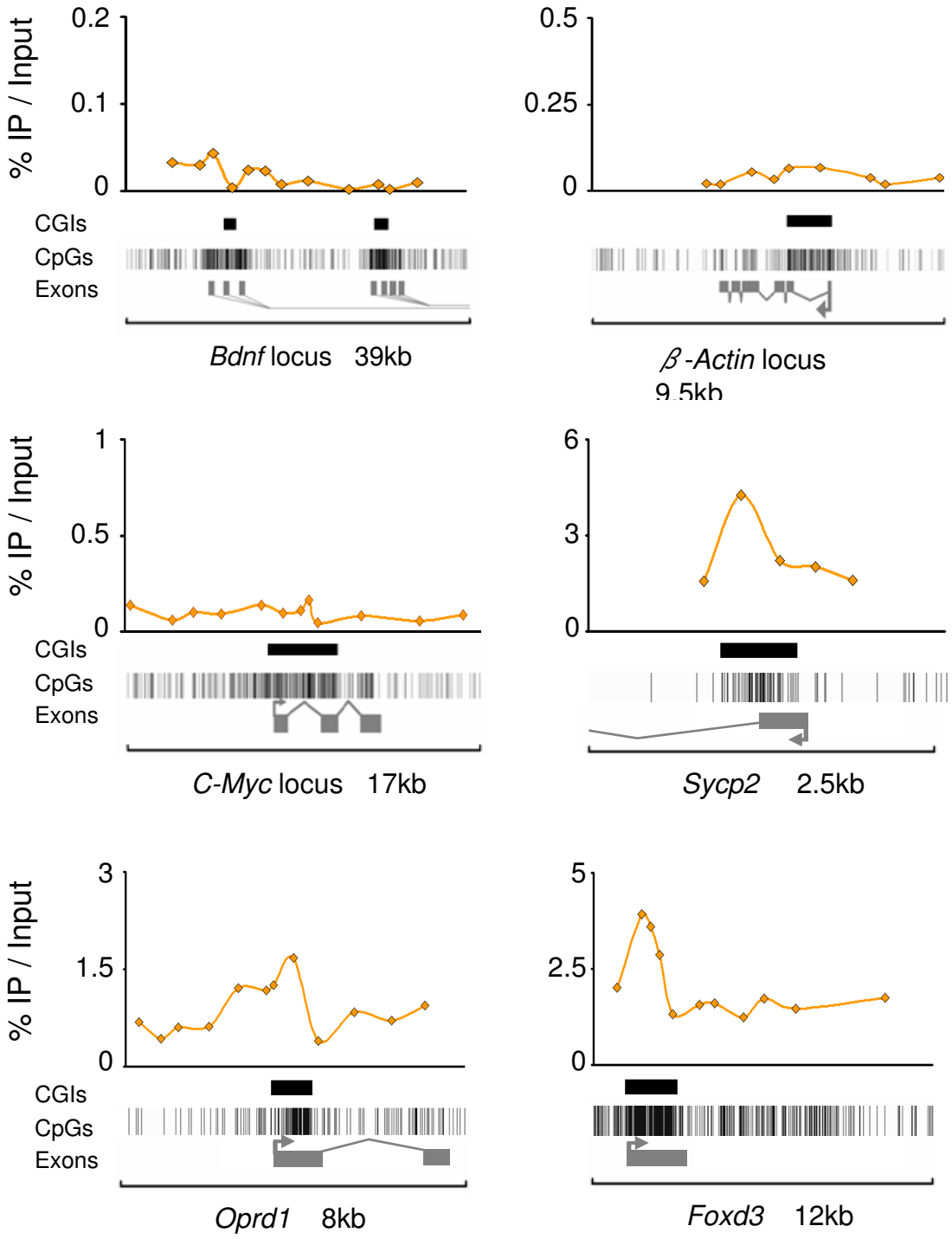

Figure 4.2-4. ChIP-PCR profiles over promoter CGIs for the histone modification H3K9me3 in mouse brain. ChIP-PCR profiling for the silencing histone modification H3K9me3 reveals limited binding over the entire *bdnf, β-actin* and *C-Myc* regions tested. In contrast, this modification is dramatically enriched over the CGIs at *Sycp2, Oprd1* and *Foxd3.*

General figure layout and abbreviations follow those outlined in figure 4.2-1

The first of these CXXC proteins tested was Mll1. Not only does this protein contain a highly conserved CXXC domain, but it also contains a SET domain responsible for the di- and tri-methylation of H3K4 residues. This protein might therefore provide an alternative link between non-methylated CpG dinucleotides and the H3K4me2/3 modification to that discussed above with Cfp1 and the Set1a/b complexes (Lee and Skalnik 2005; Lee, Tate et al. 2007). There is also some evidence to argue that the Cfp1 protein directly interacts with the MLL1 protein (Ansari, Mishra et al. 2008). ChIP profiling of the Mll1 protein over the 7 candidate CGIs revealed similar binding patterns as Cfp1. The *bdnf* locus contains peaks of enrichment for the protein which corresponded to the CGIs with 4.3 and 7.5 fold enrichment relative to the flanking regions (Figure 4.2-5). Similar patterns were seen over the CGIs at *Actb* and *C-Myc* with peaks of MLL1 enriched 2.4 and 2.9 fold relative to the flanks.

Similar to Cfp1, Mll1 is not enriched at the methylated island *Sycp2* raising the possibility that Mll1, like Cfp1, may be acting specifically as a non methyl binding protein *in vivo*. Again Mll1 follows similar patterns to that of Cfp1 over the final two CGIs of *oprd1* and *foxd3*. It is not known whether Cfp1 and Mll1 peak over the CGI regions by co-localisation to such loci or whether or not a genuine interaction exists although the findings by Ansari and colleagues would argue for the latter (Ansari, Mishra et al. 2008)

The protein Kdm2a is another member of the CXXC protein family. This enzyme, responsible for the demethylation of H3K36me3 histone tails, was also enriched over CGIs (Figure 4.1-1, B). Profiling of Kdm2a across the *Bdnf* locus revealed enrichment over the two CGIs, although with less discrete peaks than Cfp1 (Figure 4.2-6). Enrichment values of 2 and 5.2 fold over these CGIs relative to the flanking regions imply that Kdm2a may also be a CpG binding protein. Further ChIP-PCR experiments were carried out in the laboratory of Dr Rob Klose with similar CGI specific binding patterns seen across the genome (Blackledge, Zhou et al. 2010). These findings along with the earlier enrichment of Kdm2a in the CGI chromatin fraction of the genome (Figure 4.1-1) support the role of this CXXC protein as a CGI specific factor.

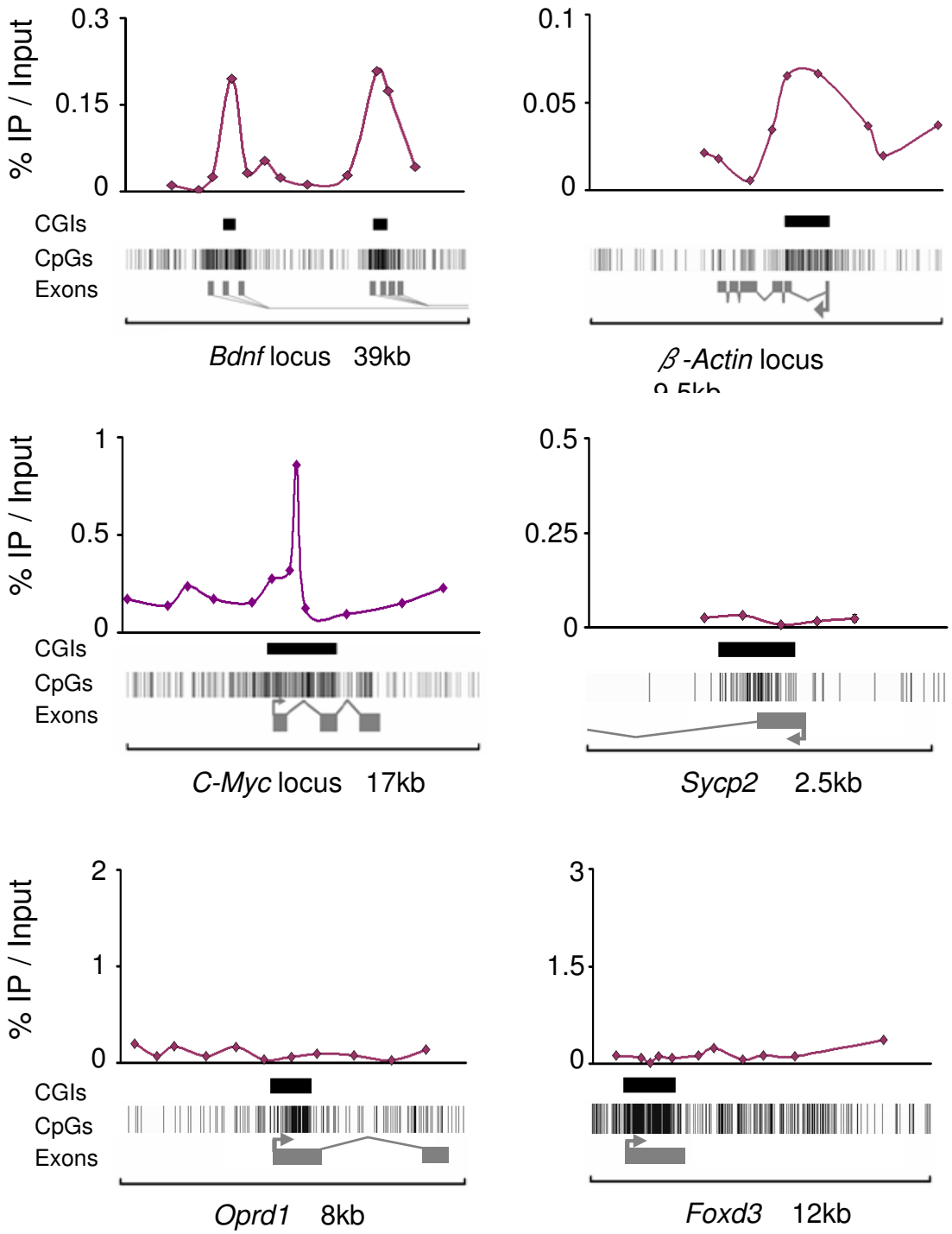

Figure 4.2-5. ChIP-PCR profiles over promoter CGIs for the CXXC protein Mll1.

ChIP-PCR profiling for the histone methyltransferase Mll1 reveals a similar pattern of binding over CGIs to that of the Cfp1 protein and H3K4me3 modification. Mll1 is seen to peak over *bdnf, β-actin* and *C-Myc* but not at *Sycp2, Oprd1* and *Foxd3.*

General figure layout and abbreviations follow those outlined in figure 4.2-1 .

Another CXXC protein of interest is Mbd1 as this protein contains both a MBD for methyl CpG binding as well as three CXXC domains for non-methyl CpG binding (Figures 1.2-4 & 1.4-1). Although this protein contains three CXXC domains only one has been shown to be capable of binding to non-methylated CpGs (Jorgensen, Ben-Porath et al. 2004). ChIP profiling for this protein over the *bdnf* locus yielded no discernable pattern of binding at or around the two CGIs. Whether or not this result represents poor immunoprecipitation or is a true representation of *in vivo* binding it is not known. Recent work using the functional CXXC domain of Mbd1 ("CXXC-3") to generate a non-methylated CGI library argues that a portion of Mbd1 is able to bind to non methylated CGIs *in vitro* (Illingworth, Kerr et al. 2008). Further work is required to better understand what is preventing full length Mbd1 from binding at CGIs.

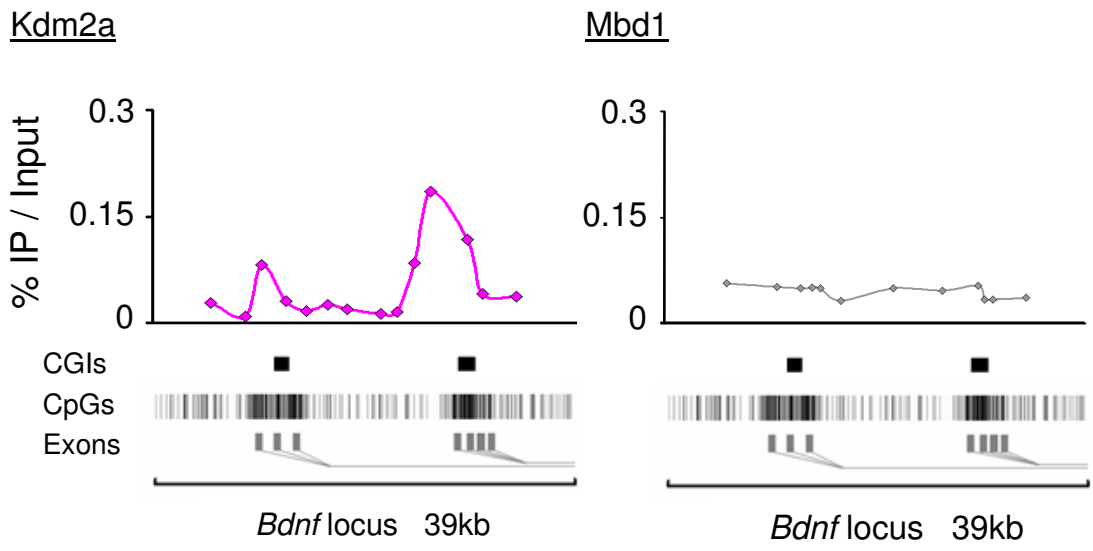

Figure 4.2-6. ChIP-PCR profiles over Bdnf promoter CGIs for the CXXC proteins Kdm2a and Mbd1. ChIP-PCR profiling of the *bdnf* locus for Kdm2a and Mbd1 give rise to dramatically different results. Although both contain CXXC domains only Kdm2a is seen to bind over the CGIs at *bdnf*. Mbd1 shows no such binding pattern, however due to the fact that the CXXC domain from this protein is used to affinity purify CGI DNA this ChIP-PCR result may be as limited by the antibody used.

General figure layout and abbreviations follow those outlined in figure 4.2-1

The CXXC domain containing protein Tet1 has been predicted to have roles in the conversion of methylcytosine to hydroxymethylcytosine (Tahiliani, Koh et al. 2009). As this process may represent an intermediate in the demethylation of DNA, the Tet1 protein is of great interest. ChIP-PCR studies over the CGIs at *bdnf* and *actb* were unable to detect any binding of the Tet1 protein. Subsequent consultation of expression data reveals that levels of Tet1 are relatively low in the brain and therefore repeating these experiments in a cell type known to contain higher levels of Tet1 (such as ES cells) may result in a different pattern of binding over CGIs. However, from these ChIP-PCR results it is impossible to determine whether or not Tet1 binds to non methylated CGIs. It cannot therefore be assumed that proteins containing this domain are intrinsically CpG island specific binding factors.

4.3 Profiling of non promoter CGIs for associated proteins

As the seven candidate CGIs interrogated above were found at promoters of known protein coding genes, one could argue that the results may be representing a promoter specific set of CGI binding proteins and modifications. PCR profiling of non-promoter associated CGIs was carried out to better understand the general protein complement.

4.3.1 Analysis of an intragenic CGI in mouse brain: exon 6 of the *atg9b* gene

The gene Autophagy related 9 homolog B (*Atg9b*) has been shown to be involved in vacuole formation in the cell (Yamada, Carson et al. 2005). Ensemble CGI predictions are reinforced by studies in the lab (Elisabeth Wachter, unpublished observations) revealing the presence of a CGI within the fourth exon of the gene. ChIP PCR was carried out over a 3.5Kb region encompassing this intragenic CGI to test for the presence of several histone modifications and CXXC proteins.

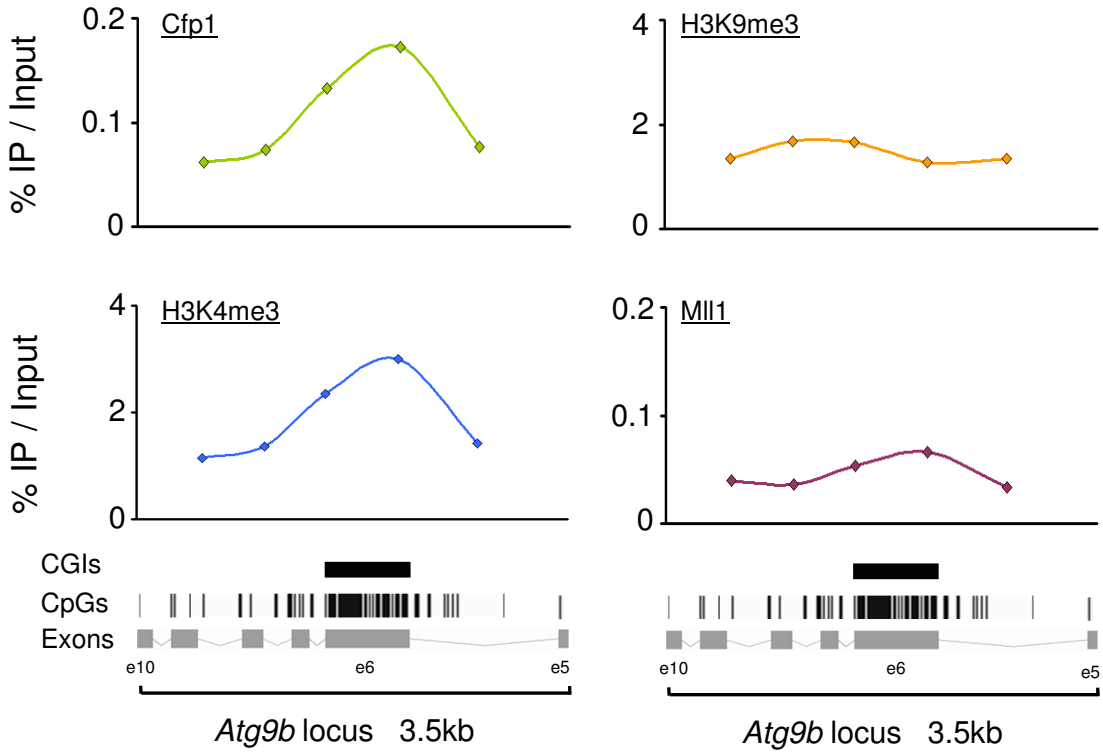

Figure 4.3-1. ChIP-PCR profiles over the intragenic CGI at *atg9b* in mouse brain. ChIP-PCR profiling for Cfp1, H3K4me3, H3K9me3 and Mll1 over the intragenic island reveals a mild enrichment of both Cfp1 and H3K4me3. Both the silencing modification ofH3K9me3 and the CXXC protein Mll1 show no such enrichment.

General figure layout and abbreviations follow those outlined in figure 4.2-1. The exon numbers are displayed underneath as "e" followed by the respective number.

Resulting PCR profiles for the CXXC protein Cfp1 showed a two fold enrichment over the exonic CGI of *Atg9b* (Figure 4.3-1). This low enrichment value is largely due to the relatively high background signal seen across this region, as the IP efficiency of the Cfp1 peak is comparable to earlier values seen over the promoter CGIs. One likely explanation for this is that the region investigated is much smaller than the earlier promoter regions tested (3.5kb here compared to 10-20kb over the promoter CGI loci). Extension of the profiled region further from this CGI so as to incorporate more flanking regions may reveal lower background values and result in a more discrete peak of binding over the CGI. Similarly, the H3K4me3 modification also contains high levels of "background" signal resulting in a low level of enrichment (2.2 fold) for this modification over the

CGI relative to flanking regions (Figure 4.3-1). Although low enrichment values are seen, both Cfp1 and H3K4me3 peak over the middle of the CGI. By contrast, the silencing modification H3K9me3 showed no preferential binding over the *Atg9b* exonic CGI region. Interestingly the other CXXC protein investigated, Mll1, did not appear to be enriched at this unmethylated CGI. This is in contrast to the earlier set of findings showing enrichment of that Mll1 at promoter annotated CGIs (Figure 4.2-5). This result may indicate promoter specific or promoter CGI specific binding of Mll1. Indeed Mll1 has been shown to have a tight association with promoter regions and the basal transcription machinery (Yokoyama, Wang et al. 2004) and as such may only be found at CGIs corresponding to promoter regions. However, due to the small region profiled and resulting background problem, no definite conclusions can be made with regards to Mll1 binding.

4.3.2 Analysis of an intergenic CGI

Around a quarter of all CGIs are found in intergenic regions, not associated with any identified gene. Such islands, termed "orphan CGIs" are likely to represent either promoters of non-annotated genes or non coding RNAs. To better understand the protein compliments at these sites and compare to promoter and intragenic CGI results, a CGI found at chromosome 5 at location 142,802,323 - 142,885,868 was investigated by ChIP-PCR for CXXC proteins and histone modifications.

Similar to the other classes of CGI investigated, the CXXC protein Cfp1 is enriched over the CGI, albeit at lower levels than the promoter CGIs (Figures 4.3-2 and 4.2-1). As was seen at the intragenic CGI at the *Atg9b* gene, these Cfp1 peaks were only enriched 2 fold over the flanking regions. Again this is possibly due to small size of the region analysed (5.1kb compared to 39kb at *bdnf*) and extension of the profiles to include more of the flanking regions may result in a reduction of the background signal. Subsequent profiling for the H3K4me3 modification reveals a similar pattern of enrichment to that of Cfp1, with an enrichment over the CGI of around 2 fold relative to the flanking regions.

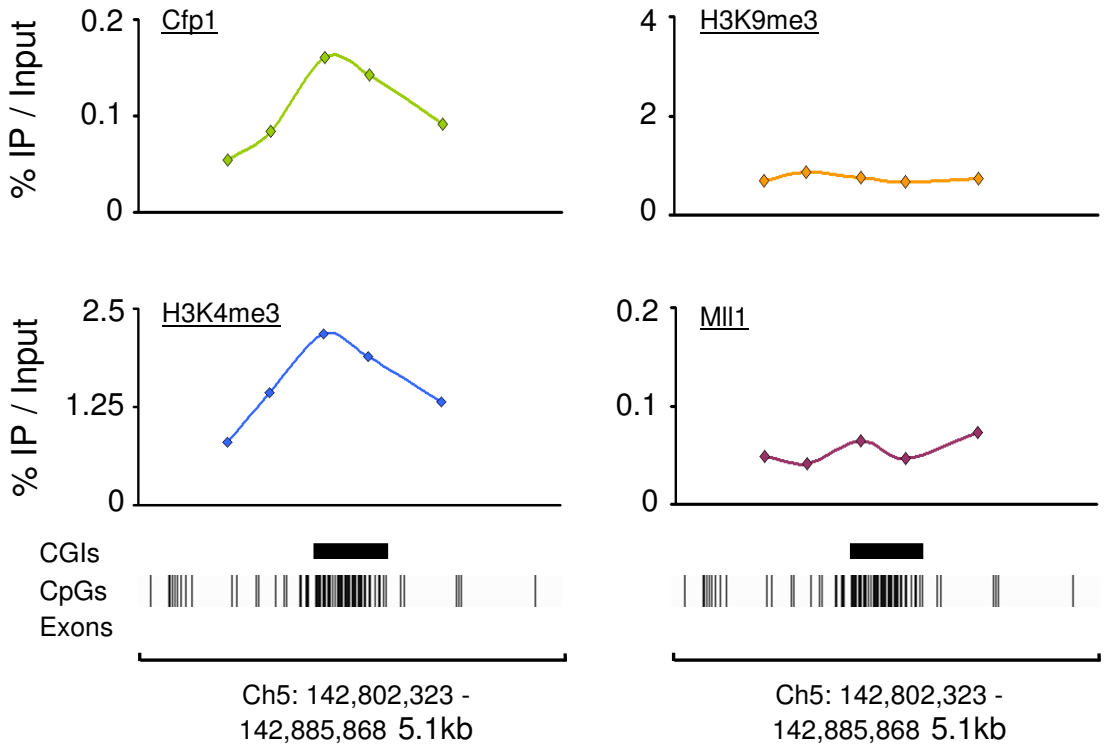

Figure 4.3-2. ChIP-PCR profiles over an intergenic CGI in mouse brain. ChIP-PCR profiling for Cfp1, H3K4me3, H3K9me3 and Mll1 over the intergenic CGI once again reveals an enrichment of both Cfp1 and H3K4me3 over the island. Again, both the silencing modification of H3K9me3 and the CXXC protein Mll1 show no such enrichment.

General figure layout and abbreviations follow those outlined in figure 4.2-1.

Analysis of H3K9me3 and Mll1 binding over the intergenic island revealed striking similarities to those seen at intragenic islands. Both the silencing modification and the CXXC protein showed no discernable binding over this site and therefore strengthen the view that these are promoter specific factors. The finding that Mll1 is not enriched at either of the non-promoter islands implies that the Cfp1/Set complex may be the H3K4 methyltransferase leading to the maintenance of the H3K4me3 modification at these CGIs. It remains to be determined whether these examples are representative of the whole genome.

4.4 Genome wide ChIP sequencing of CGI binding factors

Recent advances in genome wide sequencing has paved the way for extremely thorough and in depth *in vivo* DNA-protein studies. The ability to sequence fragments of DNA isolated by chromatin immunoprecipitation (ChIP-Sequencing or "Chip-Seq") allows high resolution protein-DNA maps to be drawn up on a genome wide scale. In contrast to earlier techniques such as ChIP-ChIP (ChIP followed by microarray hybridisation), ChIP Seq gives a higher resolution of sequencing from as little as 10ng starting material. This in turn has led to a great advancement in our understanding of the binding sites and therefore potential roles for proteins and histone modifications within the cell (Barski, Cuddapah et al. 2007; Robertson, Hirst et al. 2007; Whiteford, Skelly et al. 2009).

Methods of whole genome sequencing differ depending on the particular company and equipment used. One of the more advanced and therefore widely used whole genome sequencing approaches is that developed by Illumina. This technology relies on the sequencing of clusters of amplified immunoprecipitated DNA to build up an image of genomic protein binding sites (Figure 4.4-1, A). One sequencing run can generate more than 20 million sequence tags of up to 36 base pairs each which allows ChIP-Seq studies to be carried out with a positional resolution of +/- 50bp (Mardis 2007). A more accurate set of protein binding sites can be determined than previous ChIP-ChIP methods in which a large number of tiling arrays would be required to gain a comparable resolution.

4.4.1 Recent genome wide protein binding studies

The advent of new sequencing technology has led to an influx of high resolution studies carried out on a genome wide level. Many of these have focused specifically on histone modifications as these proteins are highly abundant and can be immunoprecipitated along with associated DNA sequences with high efficiency. The profiles of several of the histone methylation modifications such as K4, K9, K36 and K27 of histone H3 as well as RNA polymerase II were produced by ChIP-Sequencing across the promoter regions in T cells (Barski, Cuddapah et al. 2007). Promoters were grouped based on the transcriptional state of the associated gene, allowing direct comparisons to be drawn between certain histone modifications and expression levels. Although the distribution of many of these modifications

had been previously determined, this study by Barski and colleagues was able to clarify these finding at a much higher resolution.

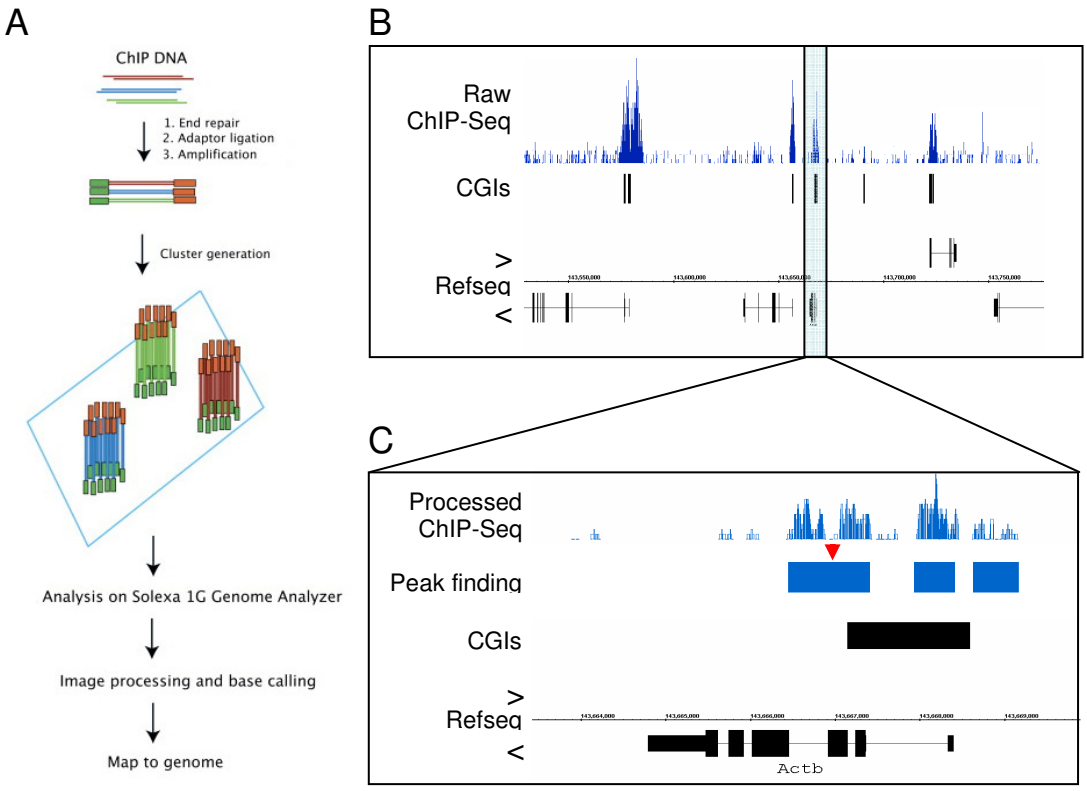

Figure 4.4-1. An overview of ChIP-Sequencing and data analysis. A. Flow chart of the steps taking immunoprecipitated DNA ("ChIP DNA") through to sequencing and mapping. Universal adaptors are added to the ChIP DNA before amplification and subsequent clustering onto plates. These plates are read by the Illumina genome analyser machine and converted into the readable and presentable format seen in **B** (figure created using the Integrated genome browser – IGB – provided by Affymetrix). At this stage the data is in a somewhat raw format (**B**). Shown is a typical display window where the ChIP seq raw data can be compared to genomic entities such as CGIs or RefSeq predicted genes. In the figure, the blue peaks correspond to protein binding patterns. The y axis represents the intensity of binding through the number of sequenced reads whilst the x axis is the position within the genome, allowing the reads to be mapped accordingly. **C**. These raw data files can be processed to remove artefacts of sequencing as well as non specific background noise. These processed files can then undergo user defined "peak finding" steps to identify true regions of protein binding (represented by the blue box under the ChIP-Seq data file). The red arrow denotes a common issue with peak finding whereby two peaks are often classed as one depending on the distance between then (the "gap") (Figure A taken from (Barski, Cuddapah et al. 2007)

This new technology has also resulted in some rather interesting findings. Traditionally RNAP II was believed to be specifically recruited to the promoters of actively transcribing genes. However, the ChIP-Seq data generated by Barski and colleagues reveal that the RNAP II protein can associate with

the transcriptional start site of nearly all promoters be they active or silent. Closer inspection reveals that higher levels of binding are seen at the promoters of highly expressed genes than those of silent.

Further sets of studies utilising the ChIP-Seq technology have been applied to a wide range of protein-DNA interactions ranging from studies determining nucleosomal positioning over promoter regions (Schones, Cui et al. 2008) to the mapping of enhancer elements (Visel, Blow et al. 2009). The latter study in particular highlights the potential of ChIP-Seq to identify the binding sites of DNA binding proteins. Therefore, ChIP-Seq can be used to thoroughly investigate the distribution of candidate CGI binding factors discovered in earlier Western blot and ChIP-PCR experiments on a genome wide scale.

4.4.2 Preparing Samples for ChIP sequencing and analysis

Although insightful, the previous ChIP-PCR studies were somewhat limited as they represented only a small fraction of the CGIs. In these experiments both the CXXC protein Cfp1 and H3K4me3 were seen to bind specifically to the CGI (see chapter 4.2-1 and 4.2.2) unless this island contained H3K27me3 modified histone tails. In order to determine whether or not this was true globally, ChIP-Seq for both Cfp1 and H3K4me3 was carried out on mouse brain tissue and compared to published H3K27me3 data sets. Although H3K4me3 has been previously sequenced through ChIP-Seq by Barski and colleagues (Santos-Rosa, Schneider et al. 2002), it was important to generate our own data sets as this would allow for the optimisation of ChIP-Seq protocols as well as to allow for the direct comparison to subsequent ChIP-Seq data sets.

In a similar approach to ChIP-PCR protocols, DNA fragments associated with the protein of interest are immunoprecipitated with a specific antibody (Figure 4.4-1 A). Universal adapters are then attached to the ends up these fragments to allow PCR amplification of ChIP material (typically 12-20 cycles). Amplified material is then clustered into groups at the Sanger institute, Cambridge, followed by sequencing-by-synthesis on analyser machines. Image files are made from these sequencing reactions which in turn are converted into DNA sequence data which can be plotted against the genome. Prior to analysis, the resulting raw sequence files undergo a round of normalisation followed

by background subtraction in order to remove non-specific noise. To these ends a set of in-house tools called GALAXY were used (found at http://bifx3.bio.ed.ac.uk:8080). The parameters for normalisation and background can be found in the Materials and Methods section.

Processed data can be viewed as a series of sequence hits over the genome using the Integrated Genome Browser (IGB) software provided by Affymetrix, which allows multiple ChIP-Seq profiles to be plotted simultaneously against the genomes of many organisms. In order to carry out statistical analysis, peaks of enrichment were defined from the raw sequencing data (Figure 4.4-1, B & C). These regions of enrichment are represented by a box marking genuine binding sites for the protein. GALAXY tools were used to convert significant peaks into blocks which can be compared between samples. In this way, sites which contain a box (referred to as a "peak") for two individual proteins could be said to be overlapping and therefore co localised. Each individual ChIP-Seq sample had a unique set of criteria for what would define a peak however some pragmatic rules apply to the length, height and gaps required to produce a box or peak. For the conditions used to determine peaks during analysis see table 2.2-1.

4.4.3 Analysis of ChIP-Sequencing data

In order to investigate the distribution of the histone modifications and Cfp1 throughout the genome, specifically with respect to the CGIs, an accurate CGI data set must be used. Predictive techniques vary wildly in content depending on the criteria selected these are not fully reliable. Fortunately a high resolution genome wide map of non methylated CpG rich DNA sequences (CGIs) was created in the Bird laboratory and would allow for the direct comparisons between ChIP-Seq results and CGI positioning (Illingworth, Gruenewald-Schneider et al. 2010). This data set was created through affinity purification using a non-methyl binding domain (a "CXXC" domain) from the Mbd1 protein. Mouse sperm DNA was used as the substrate due to the fact that resulting data sets should include all non-methylated islands. Bound DNA was eluted off the CXXC columns using salt washes and resulting fragments adapter ligated and amplified prior to ChIP-sequencing. Peak finding on this sequenced data set allowed direct comparisons and statistical analysis for overlap with later ChIP-Seq results for proteins and histone modification of interest.

Peak finding analysis on the sequenced data gave 13,348 islands, of which around 50% were found to be 5' promoter associated (Illingworth, Gruenewald-Schneider et al. 2010; Thomson, Skene et al. 2010). The remaining 50% were found to be distributed between 3' ends of genes and at intergenic and intragenic loci. Whether or not these loci are actually the promoters of non-annotated genes or non-coding RNA is as yet unknown.

i) H3K4me3

The H3K4me3 Chip-Seq was carried out and the data was analysed as outlined above. Sequence reads were plotted against the mouse genome (Ensemble mm9 build: http://www.ensembl.org/Mus_musculus/Info/Index) revealing a high coincidence of H3K4me3 enrichment over CGIs. Previous studies have shown by ChIP sequencing that the H3K4me3 mark is found enriched over promoter regions with a greater level found over active promoters than at silenced counterparts (Barski, Cuddapah et al. 2007). The data set produced here reinforces these findings as well as linking this histone modification to CGIs. The raw data was then processed so that further statistical analysis could be applied. Statistically significant H3K4me3 peaks were determined through the use of user defined parameters and were validated through comparison to the earlier ChIP-PCR results (Table 2.2-1). Once processed, the H3K4me3 data set gave 13,663 unique peaks over the entire genome. When restricted to the CGI positive data set (13,348 sites) there were 10,440 occurrences of overlap (78.2%) between the histone mark and CGIs. This reinforces the earlier ChIP-PCR and western blot studies regarding the enrichment of K4me3 modified histone H3 tails at CGI loci (Figures 4.1-1 & 4.2-2).

ii) Cfp1

Cfp1 ChIP was carried out by both myself and Pete Skene of the Bird lab using the Cfp1 Santa-Cruz H120 antibody and sequencing adapters ligated onto the purified DNA fragments by Robert Illingworth in the lab prior to sequencing by the SANGER institute. The resulting Cfp1 sequencing data contained greater levels of background "noise" than the histone H3K4me3 ChIP-Seq data set, largely due to the low IP efficiency (~0.5%) and enrichment values (2-5 fold over flanking regions) of

the Cfp1 antibody. The raw data was normalised as before (Table 2.2-1) resulting in a processed Cfp1 data set (Figure 4.4-2). Due to the low IP values seen for this antibody, two individual ChIP-Seq experiments were run in parallel and the data combined into one file. Combining data in this way should lead to a greater signal to noise ratio assuming the background noise effect is random and binding specific. The resulting combined data file was then subjected to peak finding parameters (Table 2.2-1) in order to determine statistically significant Cfp1 binding sites.

Peak finding revealed that the earlier correlation between Cfp1 and non-methylated CGIs holds true across the entirety of the genome. Of predicted CGIs, 81.5% contain a significant enrichment of Cfp1 over their length. Direct comparison of these Cfp1 positive islands with peaks of H3K4me3 confirms the co-localisation seen in the candidate ChIP-PCR studies. In total the majority (92.5%) of CGIs which are positive for Cfp1 co-localise to peaks of H3K4me3 (Figure 4.4-2).

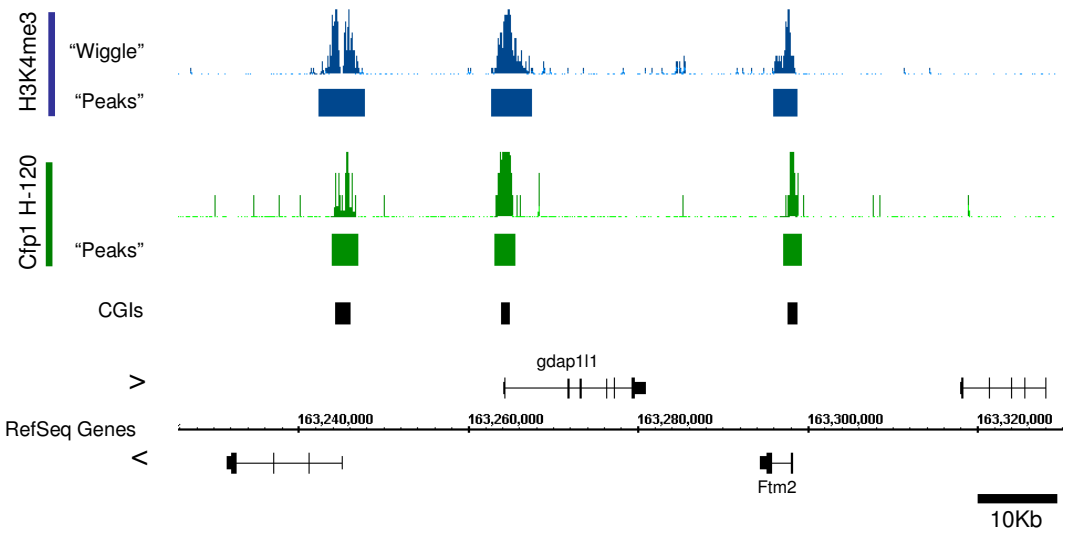

Figure 4.4-2. Raw and processed data sets for H3K4me3 and Cfp1 ChIP-Seq in mouse brain. Wiggle files and processed peak finding data sets are plotted against the mouse mm9 build using the IGB programme. "Wiggle" files represent the raw sequencing data with the height of the signal corresponding to the depth of sequencing of IP'd fragments. "Peak" data sets represent user defined regions of significant binding from the processed wiggle files. Peaks for H3K4me3 and Cfp1 were seen to overlap with predicted CGIs.

In order to ensure that this Cfp1 data set is indeed accurate, chromatin immunoprecipitation was repeated with a second antibody donated by Dr David Skalnik at the Harvard medical school. Although this antibody appears to be more specific in its binding the efficiency of immunoprecipitation is ten fold lower than the original H-120 data set resulting in higher levels of background noise. To reduce this problem samples were sequenced twice from two independent immunoprecipitations prior to combination into one data file. This had the desired effect of increasing the signal to noise ratio and allowed peak finding to be carried out with greater confidence. Although this second Cfp1 data set closely mimics the H-120 set of ChIP-Seq experiments (Figure 4.4-3) it lacks 9.2% of the Cfp1 positive islands identified in H-120 results (Figure 4.4-3, B left panel). This is likely due to both an increased specificity imposed by the antibody (as seen by the greater signal to noise ratio of peaks) and a low IP efficiency. Overall, the Cfp1 data set generated using the Skalnik antibody overlap with the H-120 data set by 96% whilst the H-120 overlaps with the Skalnik set by 85%. Combining both the H-120 and Skalnik Cfp1 data sets and scoring only for an identifiable peak in both data sets drops the percentage of Cfp1 positive islands at CGIs from 81.5% in the H-120 data set alone to 69.5% in the combined set. Comparing the Cfp1 bound islands (from both individual data sets as well as the combined set) to H3K4me3 peaks reveals that the tight correlation persists regardless of the Cfp1 antibody (Figure 4.4-3, B right panel). 96% of Skalnik Cfp1 positive islands are H3K4me3 positive which compares to the 92.5% of H-120 Cfp1 positive islands seen to contain the active histone modification. Furthermore, the combined Cfp1 data set reveals nearly 97% overlap with H3K4me3. Although the Skalnik Cfp1 antibody appears to miss many of the Cfp1 peaks identified in the previous H-120 ChIP-Seq experiments, this second ChIP-Seq experiment increases the confidence that Cfp1 is a CGI binding factor and that the protein is tightly linked to regions of H3K4me3 at CGIs genome wide. Due to the low IP efficiency associated with the Skalnik ChIP-Seq experiment, the H-120 Cfp1 data set was used in subsequent experiments.

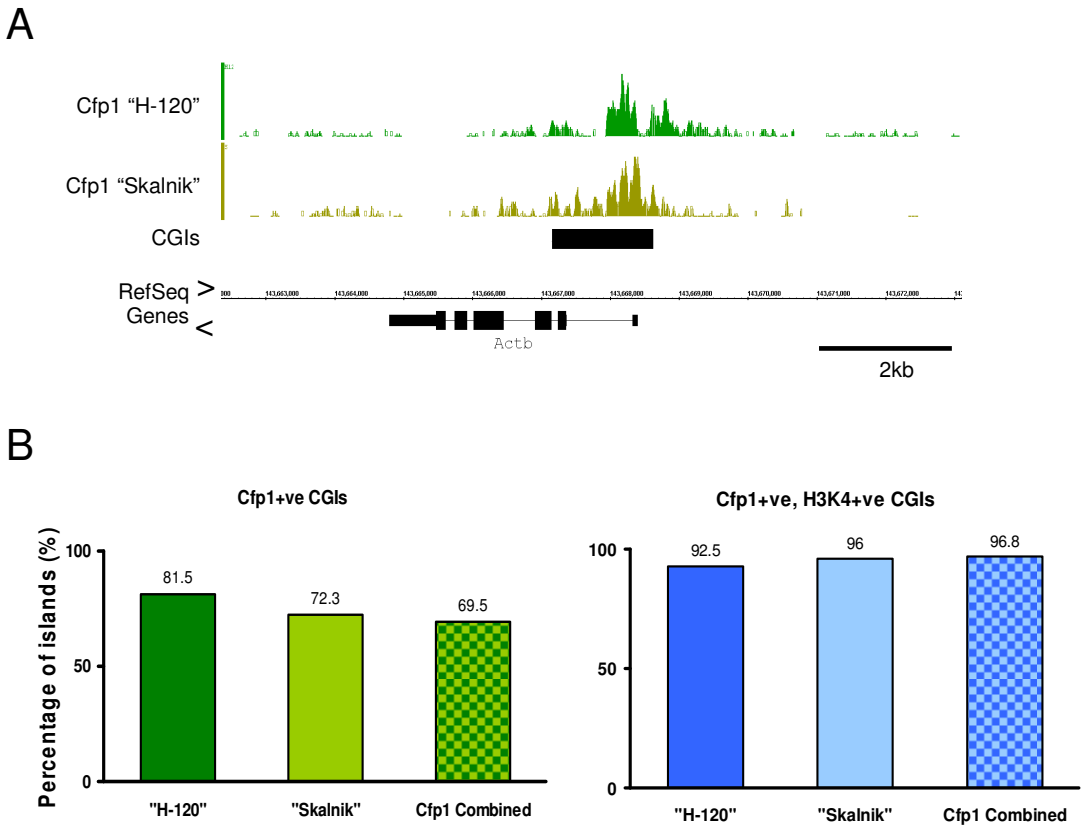

Figure 4.4-3. Comparison of two independent Cfp1 ChIP-Seq data sets. **A.** Commercial Cfp1 "H-120" antibody produced data set (upper tract, dark green) reveals a highly similar pattern of enrichment as the "Skalnik" Cfp1 antibody (lower tract, light green) over the *β-Actin* CGI. **B.** Genome wide statistics for the two Cfp1 antibodies. The left panel represents the percentage of all CGIs that are positive for either, or both ("combined") Cfp1 antibodies. The Skalnik antibody is less efficient at pulling down CGI fragments and as such combining the two data sets results in around 10% fewer CGIs than the commercial H-120 alone. The right panel represents the percentage of these Cfp1+ CGIs that are also H3K4me3 positive. In all cases the association of H3K4me3 with the Cfp1 antibodies is tight. Combining the two Cfp1 data sets gives rise to around a 97% association with H3K4me3.

iii) H3K27me3

The earlier ChIP-PCR studies found that not all non-methylated CGIs contain the CXXC protein Cfp1 and the active histone modification H3K4me3 (chapters 4.2.1 to 4.2.3). Two such islands (*Oprd1* and *foxd3*) were instead seen to contain an enrichment of the H3K27me3 modification across the length of

the CGI. The ChIP-Seq data reveals that as a proportion (18.5%) of CGIs lack Cfp1 binding (determined by the H-120 Cfp1 data set) and that these also appear to lack the H3K4me3 modification (17.3% of CGIs lack both). To test whether or not these islands are enriched the H3K27me3 modification identified in the earlier ChIP-PCR experiments, the data sets generated for H3K4me3 and Cfp1 were directly compared to previously published H3K27me3 ChIP-Seq data sets on whole mouse brain tissue (Mikkelsen, Hanna et al. 2008). As the raw data was not available, processed "peak finding" data sets were compared to our ChIP-Seq data for further analysis. Although this hinders direct comparisons between our data sets to the previously published sets it still allows for some general conclusions to be drawn.

Analysis of this data reveals that enrichment of H3K27me3 occurs over larger domains than the histone modification H3K4me3 in mouse brain, with around 50% of these domains corresponding to defined CGIs. Direct comparisons to the H3K4me3 and Cfp1 data sets reveal that the patterns seen in the earlier candidate ChIP-PCR experiments are the same genome wide (Figure 4.4-4). Typically islands which lacked the H3K4me3 modification (2908 islands) also lacked Cfp1 (2704 islands). The majority (63%) of these H3K4me3 negative, Cfp1 negative islands were found to be enriched in the H3K27me3 modification (1697 of 2707 Cfp1-ve/H3K4-ve islands). The functional significance of this is unknown, but it will be interesting to test whether of not the H3K27me3 modification itself is refractory to Cfp1 binding or the presence of H3K4me3 modifications, or if this mark is a secondary event in silencing and that is simply associated with silent regions of chromatin.

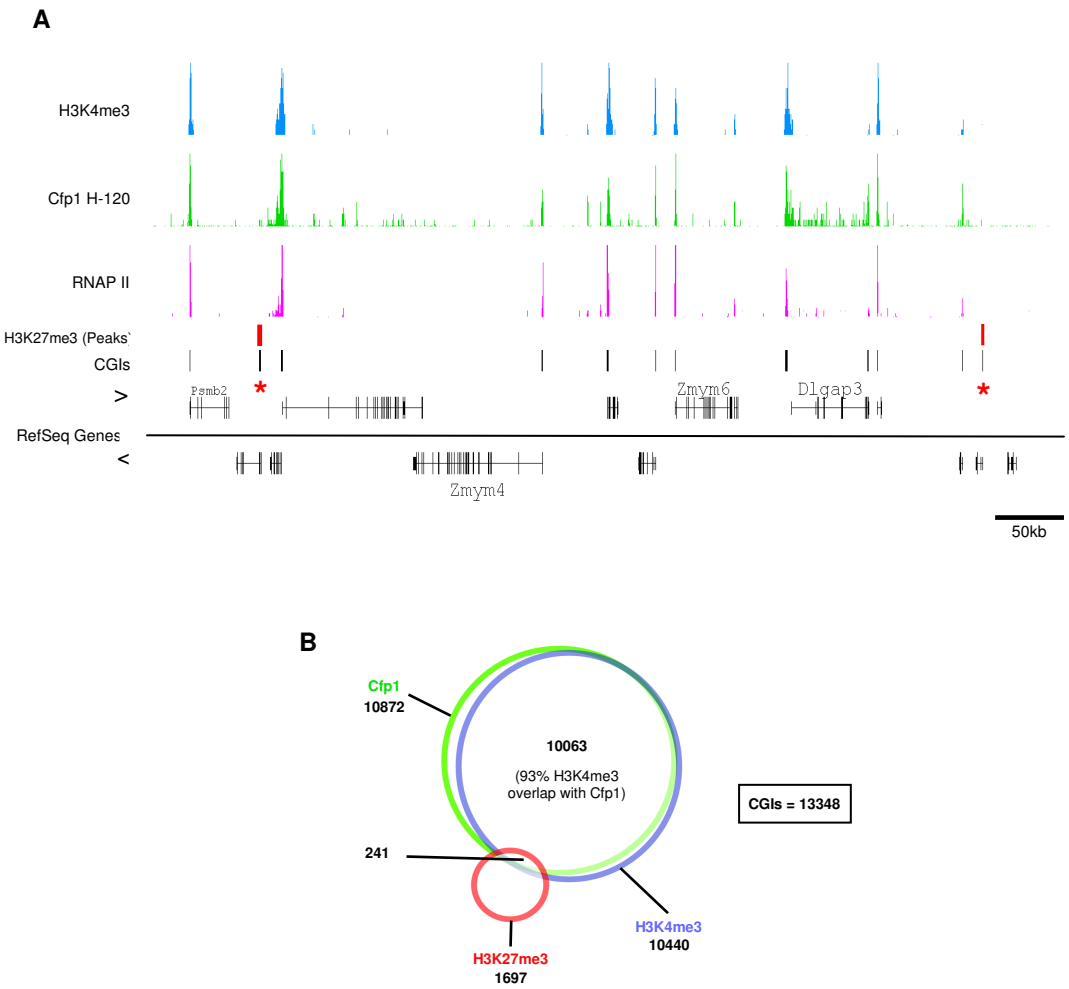

Figure 4.4-4. H3K27me3 peaks are found over CGIs which lack Cfp1 and H3K4me3. A. IGB alignment of H3K4me3 (blue), Cfp1 (green), RNAP II (pink) and H3K27me3 peaks (red boxes). Typically CGIs co-localise to peaks of H3K4me3, Cfp1 and RNAP II but not H3K27me3 peaks. H3K27me3 positive islands are indicated by the red asterisks and lack all other marks. **B.** Venn diagram representing the overlap of the H3K4me3, Cfp1 and H3K27me3 peaks at islands. Once again it is apparent that the H3K4me3 and Cfp1 peaks are tightly co-localised over islands whilst the H3K27me3 modification is discrete. Numbers indicate numbers of peaks in each group. Total number of CGIs is shown in the box.

iv) RNAPII

One of the more prominent hypotheses for the initiation and maintenance of CGIs is that transcription and the associated transcriptional machinery may either protect the DNA from methylation or facilitate the recruitment of certain histone modifying enzymes resulting in a decompacted chromatin structure (Macleod, Charlton et al. 1994; Guenther, Levine et al. 2007). Studies in which the Sp1

transcription factor binding site was mutated over the *aprt* promoter resulted in loss of transcription and caused subsequent methylation of the CpG Island (Macleod, Charlton et al. 1994). As CGIs are often found at the 5' ends of genes the process of transcription itself may be maintaining CGI state. To investigate the associations between CGI binding proteins, histone modifications and transcription, RNAPII occupancy must be investigated over CGIs. RNAP II ChIP-Seq was carried out by Pete Skene in the Bird laboratory using the non phosphorylated tail form of RNAP II. This form of the protein will be representative of recruitment of the transcriptional machinery but not of initiation or transcriptional elongation. These RNAP II peaks were found to co localise to 74.6% of all CGIs. Since only a fraction of these are expressed RNAP II appears to be pre-bound at the promoters of many genes and not simply recruited for transcription as originally thought (Figure 4.4-5,A). Direct comparisons between this data set and the H-120 Cfp1 and H3K4me3 data sets reveal an 88% overlap between all three marks (Figure 4.4-5, B). However this analysis reveals a subset of CGIs which contain both Cfp1 and H3K4me3 in the absence of RNAP II (7% of Cfp1+/H3K4me3+ islands) leading to the possibility that the non-methylated CpG sequence itself, and not transcription or the transcriptional machinery, results in peaks of H3K4me3 through the CXXC protein Cfp1. Alternatively, this finding can be explained if the ChIP-Seq experiment was missing lower affinity RNAPII binding sites. Further work on the links between RNAP II and CGI state is required to fully understand the role of this protein at CGIs (see chapter 6.4)

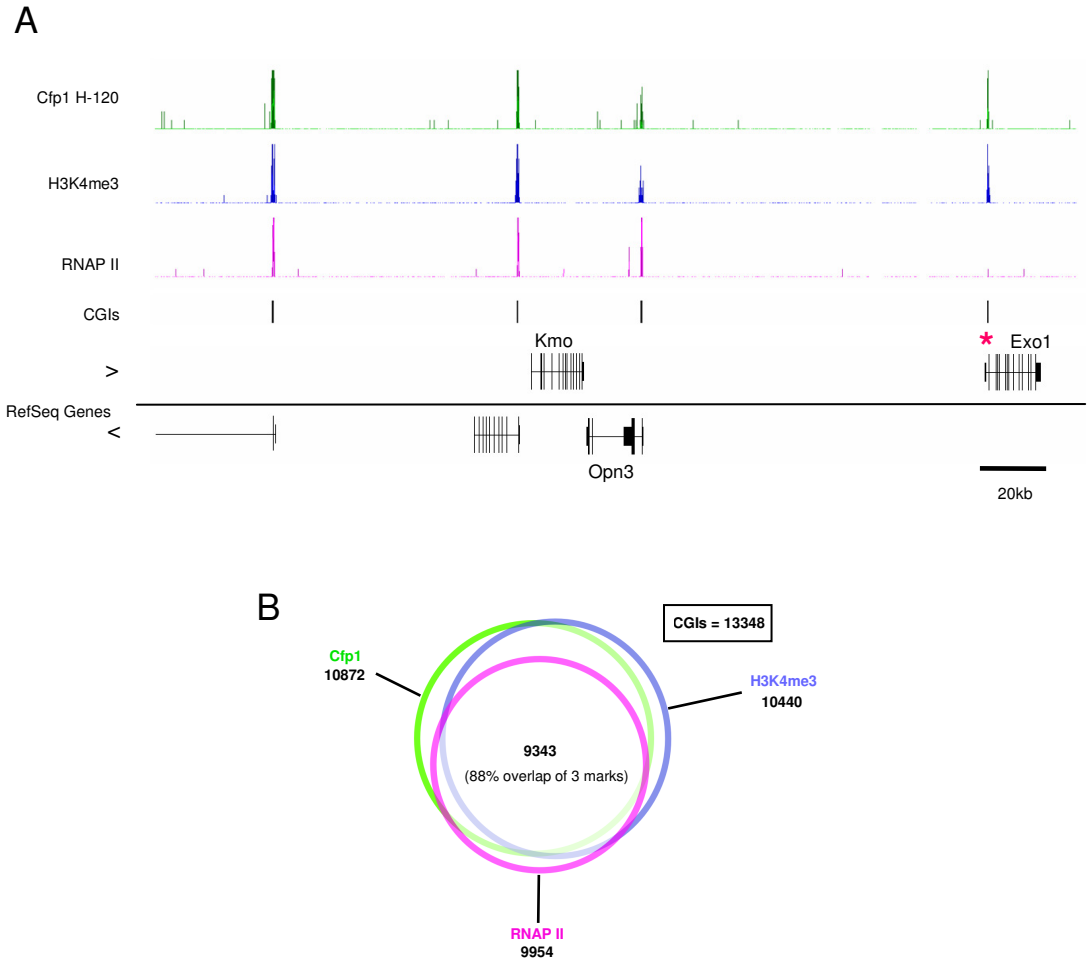

Figure 4.4-5. RNAP II ChIP-Seq data aligns to the majority of CGIs. A. IGB alignment of Cfp1 (green), H3K4me3 (blue) and non phosphorylated RNAP II (pink) along with CGIs and Refseq annotated genes. RNAP II rends to also be found associated with CGIs, however in a minority (7%) of cases is not at H3K4me3+ Cfp1+ islands (denoted by pink asterisk). **B.** Genome wide statistical analysis represented in a Venn diagram. All three proteins and modifications overlap at 88% of islands however discrete sets of CGIs exist for each. The number of peaks is represented next to each group. The total number of CGIs displayed in the box.

4.4.4 Conclusions from ChIP sequencing

The genome wide sequencing by ChIP-Seq on mouse brain material has allowed for a more expansive and high resolution set of results than the earlier candidate ChIP-PCR based experiments. However these two types of experiment give similar findings when compared directly (*β-actin* Figures 4.2-1, 4.2-2 and 4.4-3). ChIP-PCR using antibodies specific for H3K4me3 and Cfp1 resulted in profiles over

β-actin locus which closely resembled those seen in the ChIP-Seq data sets, albeit at a far lower resolution.

With regards to the findings, Both Cfp1 and H3K4me3 are found to be enriched over non-methylated CGIs, with 81.5% of islands containing Cfp1 and 78.2% the active histone modification. Furthermore the peaks of Cfp1 largely co localise to H3K4me3 peaks over CpG Islands, with a 93% overlap between the two data sets. In contrast the silencing modification H3K27me3 is generally only found at non-methylated CGIs lacking both of these proteins. As such one could argue that H3K27me3 modification is refractory to the binding of Cfp1. Alternatively this modification may be a secondary effect of silenced chromatin which is itself unable to contain either the H3K4me3 modification of the CXXC protein Cfp1.

Comparison of promoters lacking CGIs to those containing them may also delineate promoter and CGI specific factors. Although no genome wide analysis was carried out, inspection of several candidate CGI-less promoter reveals no H3K4me3, Cfp1, H3K27me3 or RNAP II signals at these loci (Figure 4.4-6, i). As promoters lacking CGIs tend to be associated with tissue specific genes, these genes may in fact be silenced in brain, although not by H3K27me3.

Finally, the finding that RNAP II also co-localises with CGIs raises the possibility that transcription or the transcriptional machinery may be responsible for the formation and maintenance of the H3K4me3 pattern at CGIs. However as a small subset of islands which contain both Cfp1 and H3K4me3 lack RNAPII, it is unclear whether or not RNAPII is required or if non-methylated CpG clusters are sufficient for the generation of peaks of H3K4me3 through the CXXC protein Cfp1.

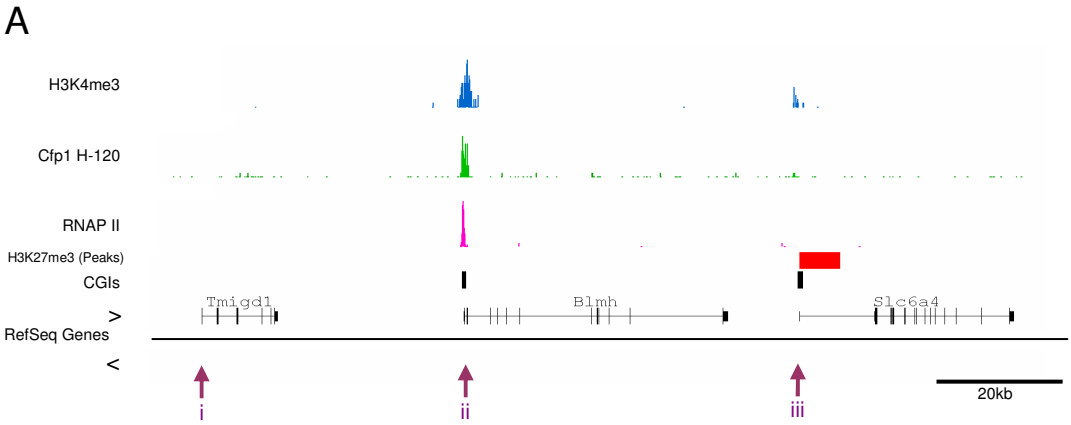

Figure 4.4-6. Comparison of a non CGI promoter to two non-methylated CGIs A. IGB alignment of H3K4me3 (blue), Cfp1 (green), RNAP II (Pink), H3K27me3 peaks (red boxes) along with CGI peaks (black boxes) and Refseq annotated genes. Group one (i) represents non-CGI promoters whilst group two (Shogren-Knaak, Ishii et al.) represents the typical CGI promoter. Group three (iii) corresponds to the H3K27me3 rich islands.

4.5 Summary of the CGI binding proteins identified

In summary, several proteins and modifications have been found enriched at CGIs through a combination of western blotting on purified CGI chromatin, ChIP-PCR over candidate CGIs and genome wide sequencing studies. Overall it appears that CpG islands are not only unique in their DNA base composition and methylation state but also harbour a distinct set of proteins and histone modifications.

Initially, the CXXC protein Cfp1 was found enriched at CpG Islands along with Kdm2a (Figure 4.1-1). Subsequent ChIP-PCR profiling for Cfp1 over several promoter and non-promoter loci also revealed enrichment of this protein specifically over CGIs (Figure 4.2-1). Notable exceptions to this rule were the methylated island of *sycp2*, and the "H3K27me3" positive islands of *Oprd1* and *Foxd3*. The fact Cfp1 is absent from the methylated *sycp2* island strengthens evidence that it may indeed be a specific non-methyl CpG binding factor. The findings of this limited candidate-style study were expanded through genome wide ChIP-sequencing of Cfp1 binding sites (Figures 4.4-2 & 4.4-4, B) which found that Cfp1 binds to 82% of CGIs mapped through CXXC Affinity purification.

The histone modifications found at CGIs were those associated with an active euchromatic state. Aside from acetylated H3 tails, H3 tails containing K4 di- and tri- methyl modifications were strongly enriched within CGI chromatin as detected by western blotting on purified CGI chromatin, candidate ChIP-PCR and ChIP-Seq. Interestingly, ChIP-PCR and ChIP-Seq both reveal that a subset of CGIs are enriched for histone H3 K27me3 modified tails. As these non-methylated CGIs were found to lack both Cfp1 and H3K4me3 this modification may be refractory for the binding of the Cpf1 protein in some way.

Taken together, these findings suggest that the CpG rich sequence found at CGIs may be functionally important in the setting up or maintenance of the local chromatin state. As Cfp1 is a CGI binding factor containing a domain with *in vitro* affinity for non methylated CpG dinucleotides and is a member of a H3K4 HMTase complex, this protein must be investigated further for its role at CGIs.

Chapter 5: CXXC proteins as non methyl CpG DNA binding proteins

5.1 Introduction

To date two of the most widely investigated CXXC proteins are the DNA methyltransferase Dnmt1 and the histone H3K4me3 methyltransferase Mll1; two proteins which have crucial roles in the establishment and maintenance of epigenetic marks. Dnmt1 is the enzyme responsible for the maintenance of methylation profiles in the mammalian genome and has been shown to harbour around a 50 fold preference for hemimethylated CpG dinucleotides (Fatemi, Hermann et al. 2001). Recent work by Pradhan and colleges has shown that the CXXC domain of Dnmt1 can bind to unmethylated CpG islands *in vitro*. Interestingly this domain appears to be vital for full enzymatic function of the Dnmt1 protein *in vivo* (Pradhan, Esteve et al. 2008). The histone H3K4 methyltransferase Mll1 has also been studied extensively not only because of its role in chromatin modification but due to the finding of chromosomal translocations which fuse the amino terminus of Mll1 to multiple loci in patients with aggressive myeloid and lymphoid leukemias (Ayton and Cleary 2001). This H3K4 HMTase has previously been shown to regulate the *Hox* genes during development with chromosomal translocations resulting in Mll1 fusions leading to *Hox* gene deregulation (Milne, Briggs et al. 2002). The distribution and binding of Mll1 in normal tissue reveals that this protein is often found associated with the RNA polymerase II complex at the 5' end of active genes (Guenther, Jenner et al. 2005). Such sites are also known to be predominantly rich in the H3K4me3 modification (Barski, Cuddapah et al. 2007) linking the presence of this CXXC protein with its resulting modification. Whether or not Mll1 is responsible for the methylation of a unique set of H3K4 tails in comparison with the other histone methyltransferases such as Set1a and Set1b is unclear.

In comparison to the aforementioned CXXC proteins, relatively little is known regarding the protein Cfp1 (see chapter 1.4.1). Early co-transfection experiments revealed that this protein acts as a transcriptional activator (Voo, Carlone et al. 2000) and later was shown to bind at the *Hox* promoter (Ansari, Mishra et al. 2008) which as stated above, is a MLL1 target promoter. Cfp1 -/- ES cells are

viable until stimulated to differentiate and show 60-80% reduction in DNA methylation with slightly elevated levels of the H3K4me3 histone modification (Lee and Skalnik 2005). More recently it was shown that Cfp1 is part of the Set1a and Set1b histone H3K4 methyltransferase complexes (Lee and Skalnik 2005; Lee, Tate et al. 2007; Ansari, Mishra et al. 2008). This could go some way to explaining the earlier ChIP PCR findings linking H3K4me3 patterns with Cfp1 occupancy (Figures 4.2-1 & 4.2-2).

Interestingly, none of these studies have linked the non methyl CpG binding of Cfp1 to CGI binding. As the earlier western blotting and ChIP experiments in brain had revealed that Cfp1 binds only at unmethylated CpG rich regions such as CGIs it is now important to further interrogate the role of this protein, if any, at CpG islands.

5.2 Bioinformatic analysis of the CXXC domain and Cfp1

To reach a greater understanding of the functions of the CXXC domain containing proteins, specifically with regards to the protein Cfp1, bioinformatic analysis was carried out. This was accomplished in two parts. Firstly, the alignment of sequences for the annotated CXXC domains in mouse was carried out to determine conservation and to allow the prediction of residues involved in non methylated CpG rich DNA binding. This analysis was followed by the alignment of the CXXC domain of Cfp1 across several organisms ranging from mouse and human to *Drosophila* and yeast.

Searching of the Uniprot protein sequence database UniProt (EMBL-EBI: www.uniprot.org) identified annotated CXXC domain-containing proteins in mouse tissues. In an attempt to better understand the amino acid residues of importance in non methyl CpG DNA binding these sequences were compared for conservation using the "Muscle" application of Jalview (www.jalview.org). These alignments highlight the high levels of cysteine conservation within the CXXC motifs across all 13 of these sequences (Figure 5.2-1) as these cysteine residues are essential in forming the CXXC zing finger structure. The carboxy- terminal end of the CXXC domain appears to vary the most across the identified sequences (denoted by the asterisk in Figure 5.2-1). Closer inspection of this region

between Cfp1 and the CXXC-3 of Mbd1 (domains which have both been shown to bind unmethylated CpGs) and the CXXC-1 and 2 of Mbd1 (shown to be unable to bind unmethylated CpG) reveal differences in sequence over this region. The fact that both Cfp1 and the third CXXC domain of Mbd1 contain a set of highly conserved residues (particularly the conservation of the residues KFGG) whilst the cxxc1 and 2 domains of Mbd1 lack these might represent a set of residues important in the binding of the CXXC domain to non-methylated CpG dinucleotides.

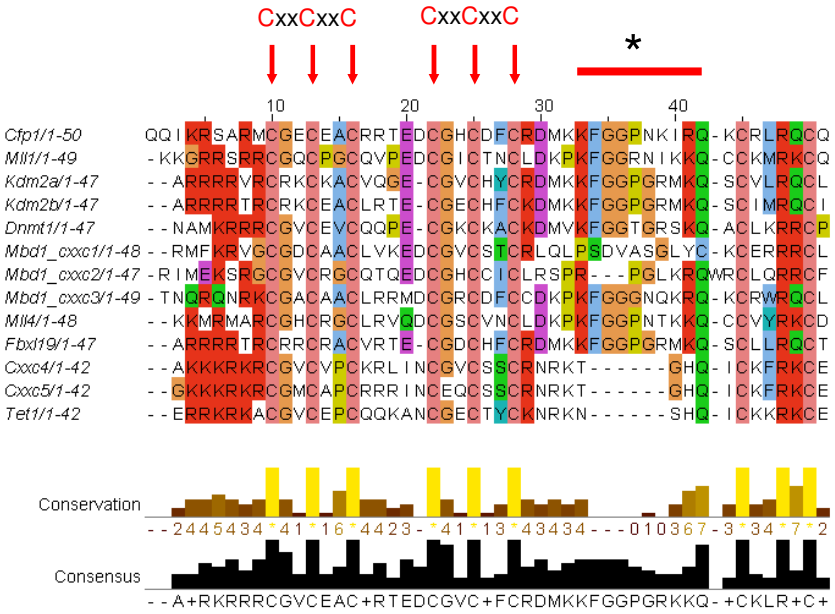

Figure 5.2-1. Alignment of identified mouse CXXC domains. CXXC domains from proteins in the mouse were found through database searches were aligned and scored for conservation. Residues are coloured as follows - red: R K, green: N S T, pink: C, magenta: E D, orange: G, gold: P, Blue: A I L M F W V. Conservation of residues is shown below with yellow bars representing high levels of conservation. A general CXXC consensus sequence is displayed below in black. Nucleotides which varied between the CXXC domains shown to bind to non-methylated CpGs to those which cannot is highlighted by the red bar and asterisk.

In an attempt to better understand the non methyl CpG binding nature of the CXXC domain of Cfp1, multiple sequence alignments across a range of organisms will give an insight into the function of the

CXXC domain along with the presence or absence of methylation. Figure 5.2-2 shows that the Cfp1 CXXC sequences of both human and mouse, as well as those found in the rat, zebrafish, *Xenopus*, the sea squirt (*Ciona*), *Strongylocentrotus* and the bee are all highly conserved. Interestingly these organisms all contain the epigenetic modification of DNA methylation. In stark contrast to these sequences are the "CXXC-like" domains found in *C.elegans, Drosophila* and *S.cerevisiae* which are all organisms in which DNA methylation is absent. Although these sequences appear to contain the highly conserved cysteine residues the rest of the sequence varies greatly (Figure 5.2-2). It therefore appears that a fully functional CXXC domain has arisen in Cfp1 during evolution, possibly to aid specific non methyl CpG protein binding in organisms containing DNA methylation.

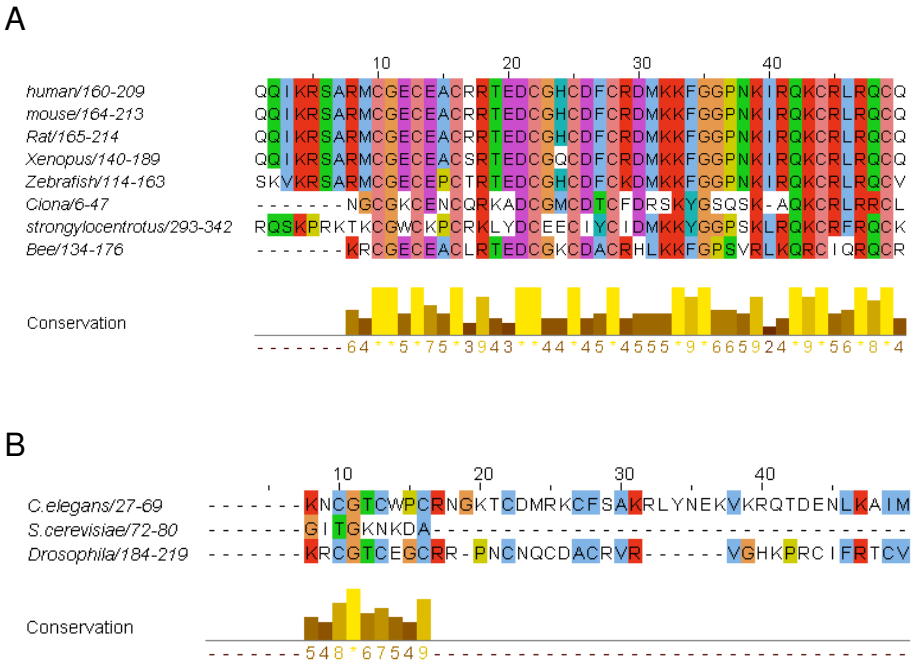

Figure 5.2-2 Alignment of the Cfp1 CXXC domain across organisms containing and lacking DNA methylation. The CXXC domains (or "CXXC like" domains) were aligned for organisms containing DNA methylation (**A**) and lacking the epigenetic modification (**B**). Residues are coloured as described in figure 5.1-1. Figure **A** reveals high levels of CXXC conservation across organisms in which DNA methylation has been identified which is in stark contrast to the CXXC-like domains found in organisms which lack DNA methylation (**B**). As the CXXC domain is important for non-methyl specific binding this may have arisen through evolution to fulfil this role.

5.3 *In vivo* verification of methyl sensitive DNA binding

In vitro experiments have previously shown that the CXXC domain of Cfp1 is only able to bind to double stranded DNA which is both rich in CpG dinucleotides and unmethylated (Voo, Carlone et al. 2000). These observations along with the multiple sequence alignments described above that Cfp1 may indeed be a non methyl CGI binding protein. This view is strengthened when one considers the results of the earlier ChIP PCR experiments in mouse brain; in particular the finding that Cfp1 is bound over the non methylated islands but absent over the methylated island at the *Sycp2* promoter (Figure 4.2-1). We therefore set out to test whether Cfp1 protein is indeed a specific non methyl DNA binding protein *in vivo*.

A simple system in which to test this further is to perform ChIP-PCR over the *Xist* gene in male and female mouse brain. *Xist* is an X linked gene found on the X chromosome of mammals which contains a strong CGI over its promoter (Norris, Patel et al. 1994). In females there are two copies of the gene, one of which is active thus non methylated and containing active histone modifications, the other is silenced and contains a methylated CGI (reviewed by (Ng, Pullirsch et al. 2007). By contrast males contain a single copy of the gene on its sole X chromosome, which is silenced and the CGI methylated (Figure 5.3-1, A). PCR profiling over this region using Cfp1 immunoprecipitated material should reveal any specificity of binding to either the methylated or non-methylated CGI.

ChIP-PCR was performed for both male and female mouse brains using antibodies against the Cfp1 protein. Figure 5.3-1 reveals that Cfp1 is bound over the CGI at the female Xist locus with 4.8 fold enrichment relative to the flanking regions. Cfp1 profiles over male brains gave no significant enrichment presumably due to the methylated island here. As the earlier studies in mouse brain revealed a potential link between the Cfp1 protein and the H3K4me3 histone modification, this mark was also profiled over the Xist locus. Patterns for H3K4me3 were similar to those for the Cfp1 protein. Female mouse brains contain a strong peak of enrichment over the non methylated CGI with 3.6 fold enrichment relative to the flanking regions. Taken with the *in vitro* CXXC domain studies (Voo, Carlone et al. 2000) this provides evidence that Cfp1 is indeed a non methyl CpG binding protein in living cells.

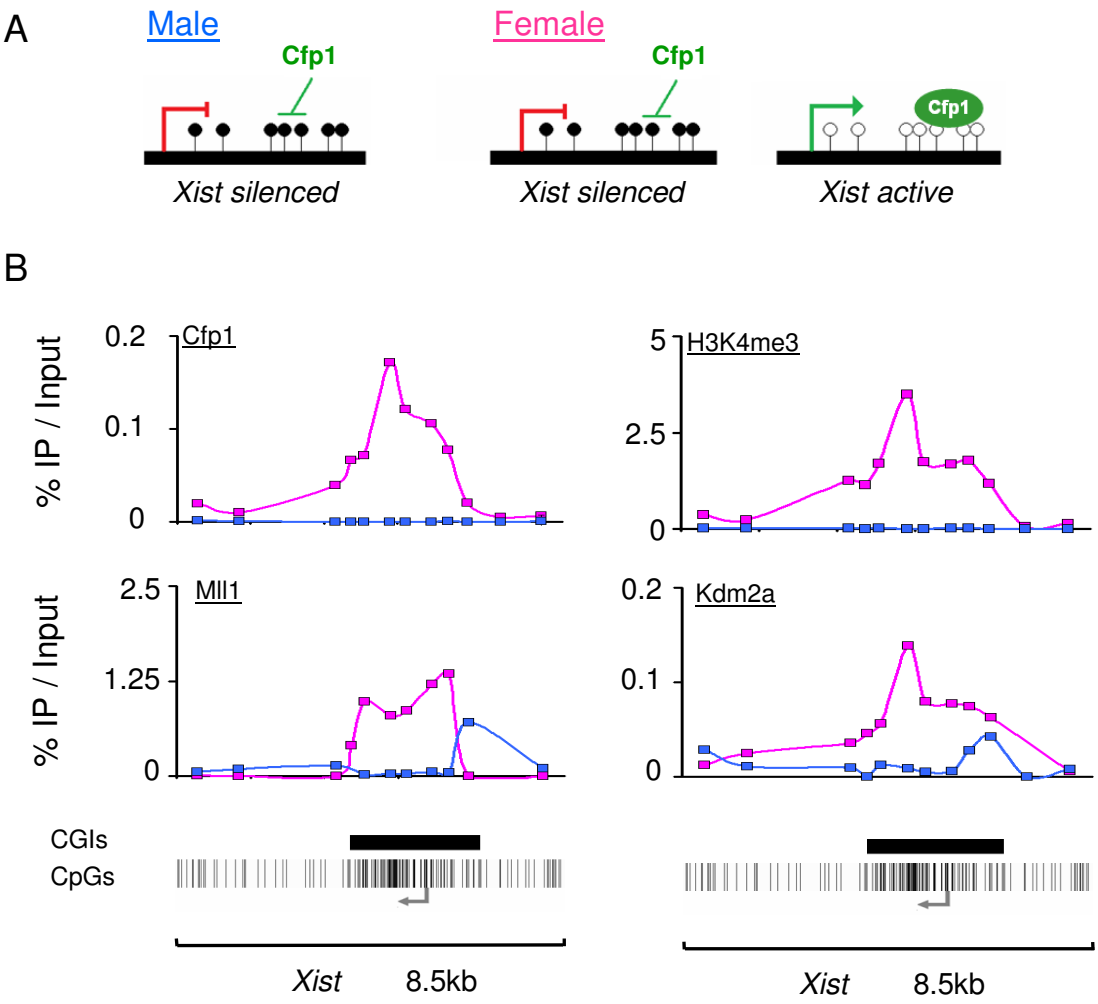

Figure 5.3-1. *In vivo* analysis revealing non-methyl CpG specific binding of Cfp1. The rationale behind the experiment is outlined in (**A**) whereby Cfp1 will only bind to the non-methylated CpGs found in female tissues. **B**. Results of the ChIP-PCR experiments over *Xist* in male (blue lines) and female (pink lines) tissues. CpG plots are represented by vertical lines under the PCR profiles whilst the CGIs are represented by black boxes. The CXXC proteins of Cfp1, Mll1 and Kdm2a are shown along with the histone modification H3K4me3.

5.4 The distribution of Cfp1 in a methylation deficient cell line

As Cfp1 has been shown to bind specifically to non-methylated CGIs in mouse brain *in vivo*, it is interesting to determine what role, if any, the abundance of CpG dinucleotides found at such CGI loci plays in the binding of this protein. Calculations were carried out on the basis that a typical human genome of $3x10^9$ base pairs contains 25,000 CGIs, each taken as around 1000 base pairs in length. As CGIs are typically C+G rich (~65%) the probability of the numbers of cytosine or guanine bases in 1000bps was calculated at 0.35 ([0.65/2]/1000). The number of CpGs in a typical 1kb island was then calculated at 105 (0.35x035x1000bp). Assuming these parameters, the human genome will contain $2.5x10^6$ CpG dinucleotides within CGIs. In contrast, a similar set of calculations predict that $3x10^7$ CpG dinucleotides occur out with such islands. This means than as few as 8% of CpG dinucleotides are in both a non-methylated state and at present within CGIs. The remaining 92% of dinucleotides are found largely methylated (70%) or non methylated (22%) at low frequency (less than 1 CpG per 10bp) throughout the genome. This argues that if the binding effect of Cfp1 is density dependant, then it can bind to approximately 8% of all the CpG dinucleotides it can encounter within the genome (those found at high densities in islands). Removal of methylation may lead to an increase in potential Cfp1 binding sites outside of CGIs which were previously CpG dense (relatively speaking) and methylated. To test the distribution of Cfp1 in such a way, *P53 -/- Dnmt1 -/-* ("P-M-") mouse embryonic fibroblasts were cultured prior to ChIP-PCR analysis. Due to the loss of Dnmt1, these cells have globally lost ~90% of their DNA methylation levels but require the subsequent removal of the P53 gene for viability (Cedar, Lande-Diner et al. 2007).

In the majority of the genome, the frequency of CpG (in a non-methylated state) is around 1 every 500 bases increasesing to around 1 every 10 bases within CGIs. This fifty fold increase in CpG density may be an important factor restricting the binding of Cfp1 to islands. In cells lacking DNA methylation (such as the P-M- cells) frequency of non-methylated CpG dinucleotides in the bulk of the genomic material increases to 1 every 100 bases whilst the frequency at islands remains the same at 1 in 10 bases, representing a ten fold increase in CpG density at the islands when compared to control cells. This change in the relative CpG densities between islands and bulk may facilitate new Cfp1 binding sites outside of CGI loci.

ChIP-PCR for Cfp1 and H3K4me3 was carried out over the *Bdnf* locus in P-M- cells with a *P53 -/-* MEF cell line used as a control. ChIP-PCR binding patterns for Cfp1 in the control cells closely follow those seen in the earlier mouse brain studies (Figure 4.2-1) with peaks of Cfp1 occurring directly over the predicted CGIs. However these Cfp1 peaks were not seen in the P-M- cell line (Figure 5.4-1). This can be explained if Cfp1 is now recruited to the newly available CpG dinucleotides outside of the CGIs which were previously methylated. Somewhat confusingly, the patterns of the histone modification H3K4me3 which peaks over the islands in the control cell line are only slightly reduced in the P-M- cell line in which Cfp1 has been completely removed. Whether or not the maintenance of H4K4me3 is facilitated by a protein complex distinct from the Cfp1-Set1 complex (such as the CXXC domain protein Mll1with H3K4 HMTase activity) is not yet known.

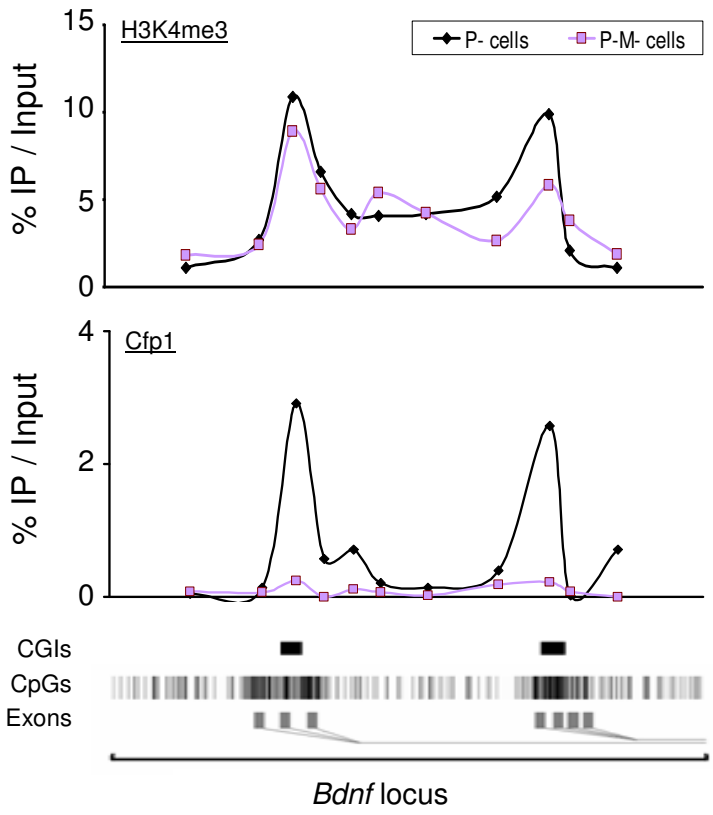

Figure 5.4-1. ChIP-PCR for H3K4me3 and Cfp1 in methylation deficient cells. ChIP PCR across the *bdnf* locus reveals a slight loss of H3K4me3 signal over the CGIs in the *P53-/- Dnmt1-/-* ("P-M-", light purple lines) cells relative to the *P53-/-* control cells ("P-", dark lines). Cfp1 by contrast is seen to be completely lost from these islands when methylation is lost globally, presumable as additionally CpG sites now compete for binding of this factor.

5.5 Summary of Cfp1 non methyl CpG DNA binding studies

From these results it can be concluded that the CXXC protein Cfp1 can bind specifically *in vivo* to high densities of non-methylated CpG dinucleotides, characteristics which are hallmarks of CpG island DNA. Taken together with the earlier western blot (chapter 4.1) and ChIP-PCR results (chapter 4.2), ChIP-PCR studies over the *Xist* locus verified this hypothesis revealing that Cfp1 is indeed a specific non methyl CpG DNA binding protein *in vivo*. Further to this, multiple sequence alignments reveal that the CXXC domain sequences of Cfp1, Mll1 and Kdm2a are highly conserved. Furthermore these sequences show significant levels of similarity to the sequence of the MBD1 CXXC3 domain which has previously been shown to bind non methyl CpGs *in vitro* (Illingworth, Kerr et al. 2008). Closer inspection of the CXXC domain of Cfp1 finds that it is highly conserved in organisms which contain DNA methylation, indicating that this domain may have evolved for the function of directing this protein to the unique regions of non methylated CpG rich DNA.

Chapter 6: Investigating the role of Cfp1 at CpG islands

6.1 Generation of Cfp1 stable knockdown cells

In an effort to better understand the role of Cfp1 at CGIs, cells deficient in the CXXC protein were generated to investigate the effects upon cell viability and growth rates as well as for any changes in the chromatin state at CGI loci. To ensure that a stable reduction of Cfp1 was achieved over a long period of time, short hairpin RNA oligonucleotides (shRNAs) specific for Cfp1, were transfected into NIH-3T3 cells. Once transcribed, the short sequences form an RNA hairpin structure which is bound by protein complexes such as DICER and converted into short interfering RNAs (siRNA). These siRNAs then bind to their complimentary sequences on the Cfp1 mRNA resulting in a cleavage event and subsequent gene silencing. As shRNA sequences remain on the transfected vector and are usually passed onto daughter cells this results in the stable and long term silencing of Cfp1

Three unique shRNA sequences averaging 19 nucleotides in length were synthesized by Oligoengine which would recognise the Cfp1 gene in mice. The resulting sequences were then cloned into a vector so that the sequences were under the control of the polymerase-III H1-RNA gene promoter which produces a small RNA transcript lacking a polyadenosine tail ("pSuper" vector). These sequences were designed to recognise the following regions on the cfp1 mRNA: 986-1005 (termed sh986), 1250-1269 (sh1250) and 1960-1979 (sh1960; see Materials and Methods section 2.1.4 for sequences). Successful transfections were selected for through puromycin resistance which was encoded in the vector. Each of the three shRNA vectors were individually transfected into NIH-3T3 cells and grown in puromycin containing media to select for successful transfectants. Additionally all three shRNA vectors were mixed and transfected to see if there was either a greater transfection efficiency or knockdown of Cfp1 when combined together to that seen with the individual transfections. Puromycin selection lead to loss of around 90% of the cells with survivors growing to form small colonies within 24 to 48 hours. Colonies were picked and grown up further to produce 40 cell lines in total (ten clones for each of the individual and mixed set of shRNA vectors).

6.1.1 Verification of Cfp1 deficient cell lines

For each transfected cell line, protein was taken for western blot analysis at the same time as RNA for quantitative PCR (qPCR) and chromatin for ChIP-PCR. As it was the simplest approach, qPCR for Cfp1 expression was carried out first to identify the levels of Cfp1 reduction in the cell lines (Figure 6.1-1). As it was believed that the cell lines most likely to be depleted in Cfp1 were those which were transfected with a mixture of all three short hairpin RNA sequences ("shMix" cells), these cell lines were tested first (Figure 6.1-1).

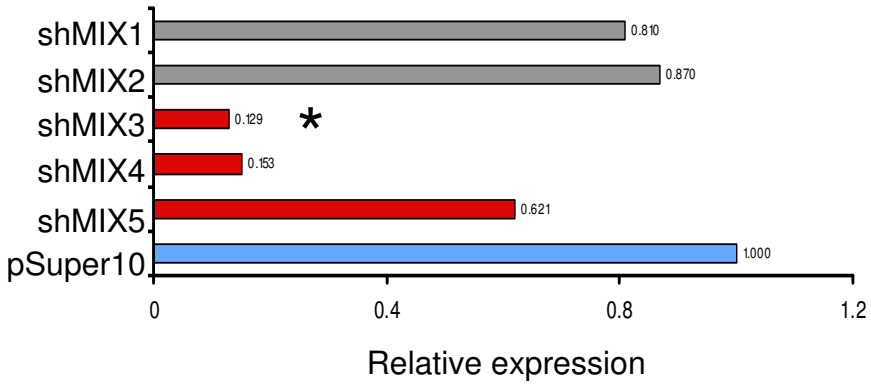

Figure 6.1-1. qPCR verification of Cfp1 levels in "mix" transfected cell lines. qPCR results for the expression of Cfp1 relative to the control cell line (PS10 vector - blue) and knockdown cells (red: successful knockdown, grey: unsuccessful knockdown). The cell line selected for subsequent ChIP-PCR studies is denoted by an asterisk. All data is normalised to the expression of GAPDH in each cell line.

Although ten cell lines were initially selected for growth after transfection with the mixture of short hairpin constructs, only five survived expansion for analysis. One possibility is that the knock down of Cfp1 in these cell lines may have been above a critical level for survival. However, this argues that the surviving cell lines are likely not efficiently reduced in Cfp1 levels. Alternatively the failure to expand may be due to the fact that these cell lines may have not taken up (or have lost) the puromycin resistance imposed through the sh*Cfp1* transfected plasmid.

Closer inspection of the surviving mixed transfected cell lines by qPCR reveal that three out of the five cell lines expanded were deficient for Cfp1 when compared to the "pSuper10" vector control cell line (Figure 6.1-1). Of these two were significantly reduced to around 13% and 15% of control Cfp1 mRNA levels. The lowest of these (termed shMIX3) was expanded and qPCR repeated a further two times resulting in an average of 87% reduction in Cfp1 levels. In order to ensure this level of knockdown was reproducible these mixed pool transfections were repeated at a later date with near identical levels of Cfp1 reduction (data not shown).

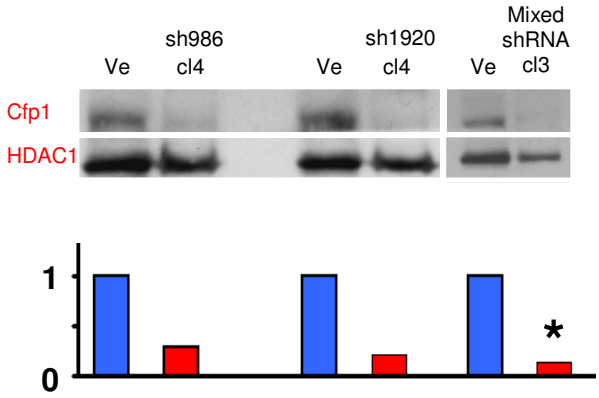

Figure 6.1-2. Western blot verification of Cfp1 levels in shRNA transfected cell lines. Results of western blots over the most Cfp1 deficient cell lines. Shown are cell lines in which individual shRNA constructs have been used (sh986 and sh1920) along with the mixed shRNA cell line previously characterised by qPCR (Figure 6.1-1) all compared directly to the vector control "pSuper10" cell line ("Ve"). HDAC1 is used as a control to ensure equal levels of protein are loaded. The mixed shRNA cell line used in subsequent ChIP-PCR experiments is highlighted by an asterisk. Graphical representation of Cfp1 levels through quantification carried out using ImageJ software on the western blot are shown directly below (blue boxes represent control vector whilst red the knockdown cell lines)

To ensure that the qPCR results measuring Cfp1 mRNA are correct, western blot analysis was carried out on the shMix3 cells to test for Cfp1 protein levels. Further to this, cell lines transfected with only one of the three shRNA constructs ("sh986" and "sh1920") were also tested for Cfp1 levels (Figure 6.1-2). Quantification of signals was carried out using ImageJ software. In concordance with the earlier qPCR results, the mix3 cell line contains around an 85% reduction in the Cfp1 protein. The 986 clone 4 cells line gave the greatest individual knockdown of Cfp1 at around 80% relative to control

cells whilst the sh1920 clone 4 had around a 75% reduction in Cfp1 levels. In conclusion NIH-3T3 cell lines were successfully knocked down for Cfp1 at the level of mRNA as well as protein in both individual and mixed transfected cell lines.

6.1.2 Growth and morphological changes associated with Cfp1 deficient cells

As Cfp1 had been successfully reduced in several cell lines, the next step was to investigate the effect that this depletion placed upon the growth rates and cellular morphologies of the NIH-3T3 cells themselves. To this end, the Cfp1 deficient cells which showed the greatest knock down ("shMIX3") were grown alongside Vector only controls ("pSuper10") and WT 3T3 cells. It was noted that the wild type NIH-3T3 and pSuper10 vector control cells required splitting almost twice as often as the Cfp1 deficient cells (Figure 6.1-3, A). Equal numbers of cells were seeded for both vector control and Cfp1 deficient cell lines and the numbers of cells counted over a period of 90 hours (for a detailed procedure see Materials and Methods section 2.2.7). When represented graphically it can be seen that the Cfp1 deficient cells grow at a lower rate than vector only cells (Figure 6.1-3, A). From these values the doubling time for vector control 3T3 cells was found to be around 13 hours which compares to around 18 hours for Cfp1 deficient cells (a 38.5% slower growth rate). Likely explanations for this lowered growth rate are that Cfp1 may be involved in cell cycle progression or that Cfp1 is required for viability in some way. This could be investigated further through FACS analysis in order to test whether or not the cells were arresting at a certain stage in the cell cycle. Interestingly Cfp1 -/- ES cells generated in the lab of Dr Skalnik are reported to have increased doubling times due to an increase in apoptosis (around 33% of all Cfp1 -/- cells (Carlone, Lee et al. 2005).

Closer inspection of the Cfp1 deficient cells reveals that they are morphologically different from the control transfected NIH-3T3 cells (Figure 6.1-3, B). Vector control cells look like their normal wild type counterparts with long flat bodies contacting neighbouring cells. This indicates that the process of transfection does not alter the structure of the cell. In contrast the Cfp1 deficient cells appeared far

more rounded in shape and clump together with reduced contacts between the cells. Therefore the removal of Cfp1 leads to an overall change in both the structure of the cell as well as the rate of growth in culture.

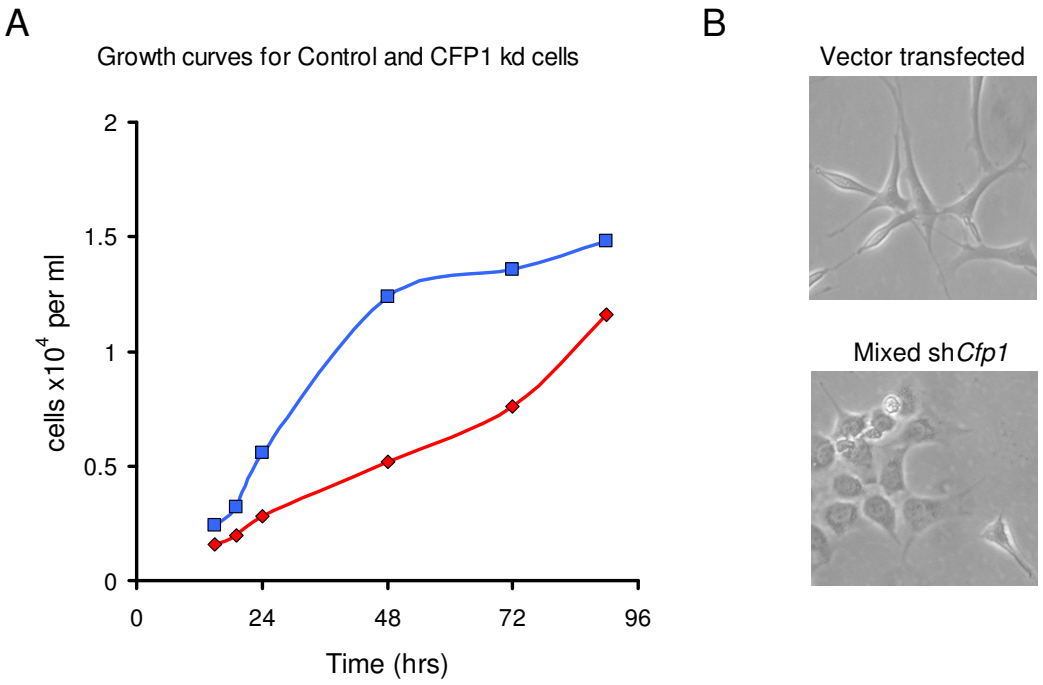

Figure 6.1-3. Sh*Cfp1* cells exhibit growth and morphological changes. Figure **A** highlights the differences in growth rates between control cells (blue line) and sh*Cfp1* mix cl3 cells (red line). Overall these Cfp1 deficient cells grow around 40% slower than the control cells. **B**. Cell morphology differs greatly between the control and Mixed sh*Cfp1* cells (original magnification x200).

6.1.3 Western blot analysis of histone modifications in Cfp1 deficient cell lines

As Cfp1 has a positive correlation with the H3K4me3 modification (chapters 4.2.1 and 4.2.2) and a negative correlation with the H3K27me3 modification in chromatin (chapter 4.2.3), global levels of these two marks were investigated through western blot analysis to compare control and Cfp1 deficient 3T3 cells. In order to assess the relative levels of global histone modifications between the cell lines, a range of concentrations of vector control cell extract was loaded alongside the sh*Cfp1* cell

lines sh986 cl4 and sh1920 cl4 and the mixed hairpin cell line (Figure 6.1-4). As the analysis was carried out using fluorescent LI-COR antibodies, precise quantification of signals was achieved.

.

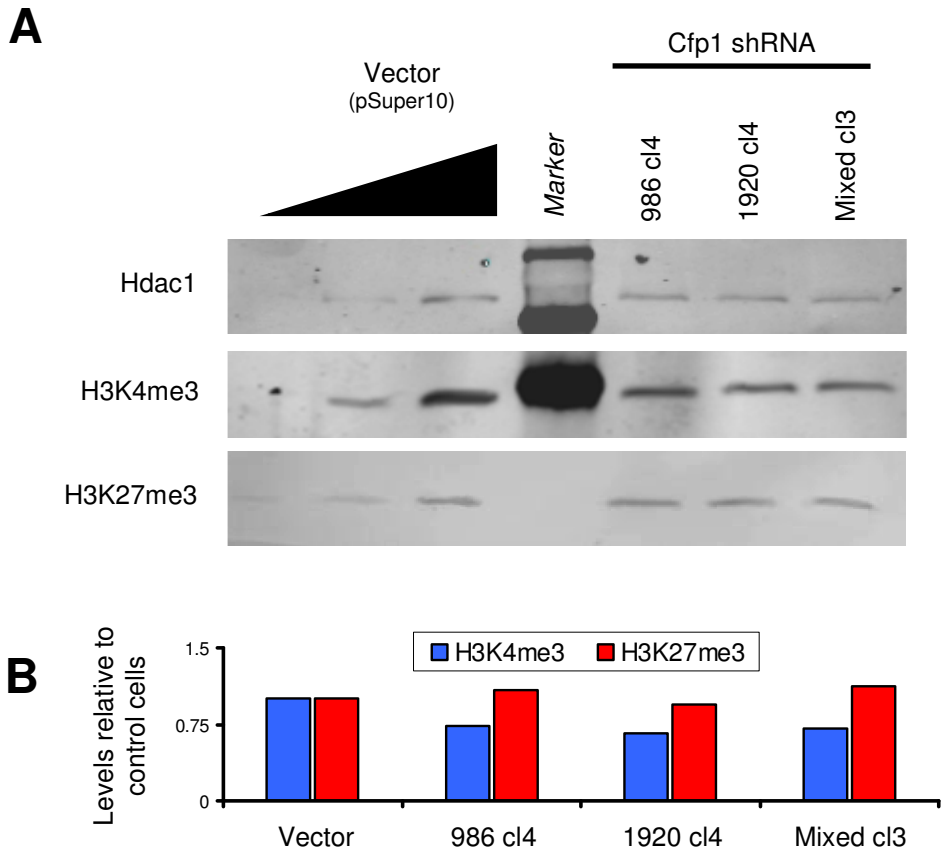

Figure 6.1-4. Global analysis of H3K4me3 and H3K27me3 in sh*Cfp1* cell lines. In **A,** LI-COR western blotting for H3K4me3, H3K27me3 and HDAC1 (as a loading control) is shown with the quantification shown underneath in **B**. a ten fold dilution range (1/10, 1/3 and equal levels) of vector whole cell extract are loaded to facilitate accurate quantification of signals. In the 986 clone 4, 1920 clone 4 and mixed clone 3 cell lines (all shown to be Cfp1 deficient – figure 5.3-1) H3K4me3 levels are slightly decreased whilst H3K27me3 remains constant. Quantification of the signals in **B** reveals that H3K4me3 levels (blue bars) fall to an average of around 70% of that seen in the vector control. H3K27me3 levels (red bars) do not appear to fluctuate.

Quantification revealed modest levels of H3K4me3 depletion in all three knockdown cell lines, averaging 70% that seen in wild type vector transfected control cells (Figure 6.1-4, B). All three cell

lines contained very similar levels of H3K4me3 (73, 66 and 71% wild type levels respectively). In contrast to the reduction in H3K4me3 levels, no such change was noted for the silencing modification H3K27me3 in each of the three cell lines.

6.2 ChIP profiling of genes using Cfp1 deficient cells

As the global levels of both H3K4me3 and H3K27me3 had been investigated in the Cfp1 deficient cell lines, ChIP-PCR profiles for these modifications along with Cfp1 were carried out to better understand any changes taking place specifically at CGIs. The same promoter CGIs that were earlier selected for analysis in the mouse brain (see chapter 4.2) were now investigated in the vector control and mixed short hairpin ("shMix cl3") cfp1 deficient cell lines (Figures 6.2-1 & 6.2-2). Furthermore, the ChIP profiles generated from the cell lines transfected with a mixture of short hairpin constructs ("shMix cl3") were compared to the profiles from the individually transfected cell lines ("sh986 cl4" and "sh1920 cl4") in order to investigate the possibility of any off target effects which can occur with shRNA silencing (see Figure 6.2-4 below).

6.2.1 Cfp1 profiles over CGIs

The Cfp1 depleted "shMix cl3" cell line was expanded prior to ChIP PCR analysis over the six CGIs previously investigated in mouse brain. As a positive control, a vector transfected cell line containing wild type levels of Cfp1 ("pSuper10") was tested along with the Cfp1 deficient cells. These control cells gave similar results to wild type NIH-3T3 cells (data not shown) with peaks of Cfp1 seen to correspond to the predicted CGIs whilst remaining low over flanking regions (blue plot, Figure 6.2-1). Typically the enrichment values of these peaks relative to flanking regions are similar to the values determined earlier in the mouse brain (Figure 4.2-1). For example the enrichment values for Cfp1 over the CGIs at *Bdnf* are 3.2 and 4.4 fold enriched in the control vector NIH-3T3 cells and 4.1 and 3.2 fold enriched in the brain (Figures 6.2-1 and 4.2-1). ChIP-PCR profiles over the same CGIs in "shMix3" cells reveal a loss of Cfp1 from these islands providing a further conformation of Cfp1depletion. Quantification of the loss of the Cfp1 signal results in a 93 and 66 fold drop over the CGIs at the *Bdnf*

locus, a 23.6 fold reduction over the island at *β-Actin*, a 41.5 fold reduction over the island at *C-Myc* with modest drops of 8.5 & 7.5 fold seen over the CGIs at the *Dlx 5/6* locus.

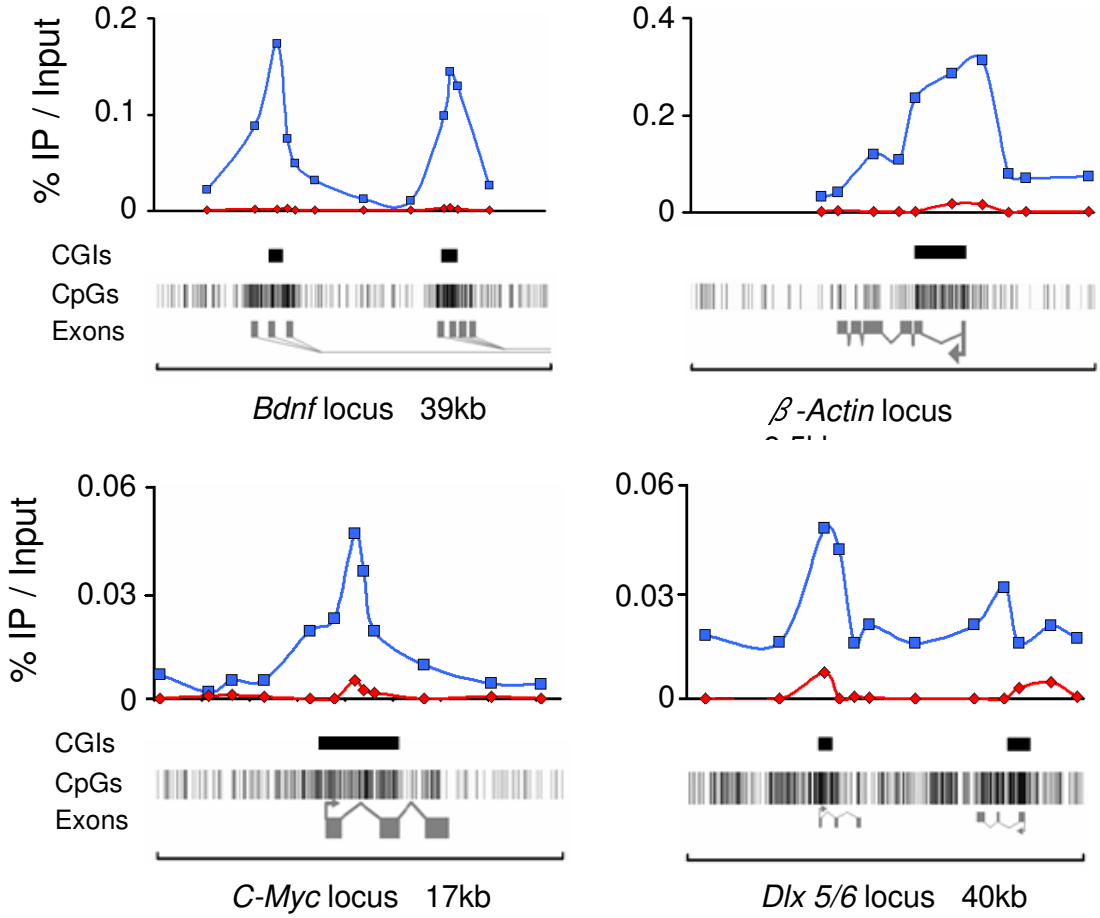

Figure 6.2-1. Cfp1 binding is dramatically reduced over CGIs in sh*Cfp1* mix3 cell lines. ChIP-PCR profiles for four loci containing 6 CGIs. pSuper10 Control 3T3 cells (blue lines) contain previously described patterns of Cfp1 binding over CGIs. These peaks of Cfp1 are lost or dramatically reduced in the Cfp1 deficient cell line "mix3" (red lines). Profile shown represents one experiment with results typical for n=3 runs.
General figure layout and abbreviations follow those outlined in figure 4.2-1.

6.2.2 H3K4me3 profiles over CGIs

Both candidate ChIP and ChIP-sequencing in mouse brain highlighted a potential link between the H3K4me3 modification and Cfp1 occupancy at CGIs. With this in mind, the patterns of H3K4me3 at CGIs were investigated in the vector control and "shMix3" Cfp1 deficient cell lines. Again the patterns of H3K4me3 in the pSuper10 vector control cells followed those seen earlier in the mouse brain (Figure 4.2-2) peaking specifically over the predicted CGIs (Figure 6.2-2). Enrichment values for the H3K4me3 modification over the CGIs were calculated at 11.4 and 10.7 over the CGIs at *Bdnf*, 12.6 fold over *Actb*, 10.3 fold over *C-myc* and 2.5 and 2.4 fold over the islands at the *Dlx5/6* locus. Although total levels of the H3K4me3 modification are only slightly reduced in the Cfp1 deficient cells (Figure 6.1-3) the association of this histone modification to CGI loci appears to be greatly reduced. Quantification of the ChIP-PCR results reveals that levels of the H3K4me3 modification over the CGI at the *bdnf* locus are 6.6 and 4.8 fold lower than those seen over the island in the control cell line. The other regions profiled also reveal a precipitous drop in the H3K4me3 signal over the predicted CGIs with patterns at *β-actin, C-myc* and *Dlx5/6* falling by 3.6, 3.7, 2.8 and 4.3 fold respectively.

From these results it is clear that depletion of Cfp1 leads to a dramatic loss of the H3K4me3 peaks normally associated with CpG islands. As eluded to earlier, Cfp1 is part of the Set1 complex responsible for setting up the H3K4me3 modification and as such may provide the enzymatic complex with a CXXC domain to aid targeting to CGIs. Depletion of Cfp1 with concomitant removal of the CXXC domain from the Set1 complex may result in the reduction of H3K4me3 patterns over islands. Any redistribution of the H3K4me3 mark throughout the genome due to lack of targeting via the CXXC domain of Cfp1 coupled with loss from islands may explain the earlier finding that global levels of H3K4me3 are only slightly reduced (Figure 6.1-4). Alternatively as Cfp1 is a core component of the Set1 complex, removal may destabilise or lead to the loss of activity of the enzymatic complex resulting in a reduced H3K4me3 signal. Studies in yeast with the Cfp1 homologue *Spp1* argue against this hypothesis as removal of this protein does not affect global levels of the Set1 protein but does result in a lower level of H3K4me3 globally (Takahashi, Lee et al. 2009). Further work on the interactions between Cfp1 and the Set1a and b complexes in both the wild type and Cfp1 deficient cells is required to better understand the H3K4me3/Cfp1 link.

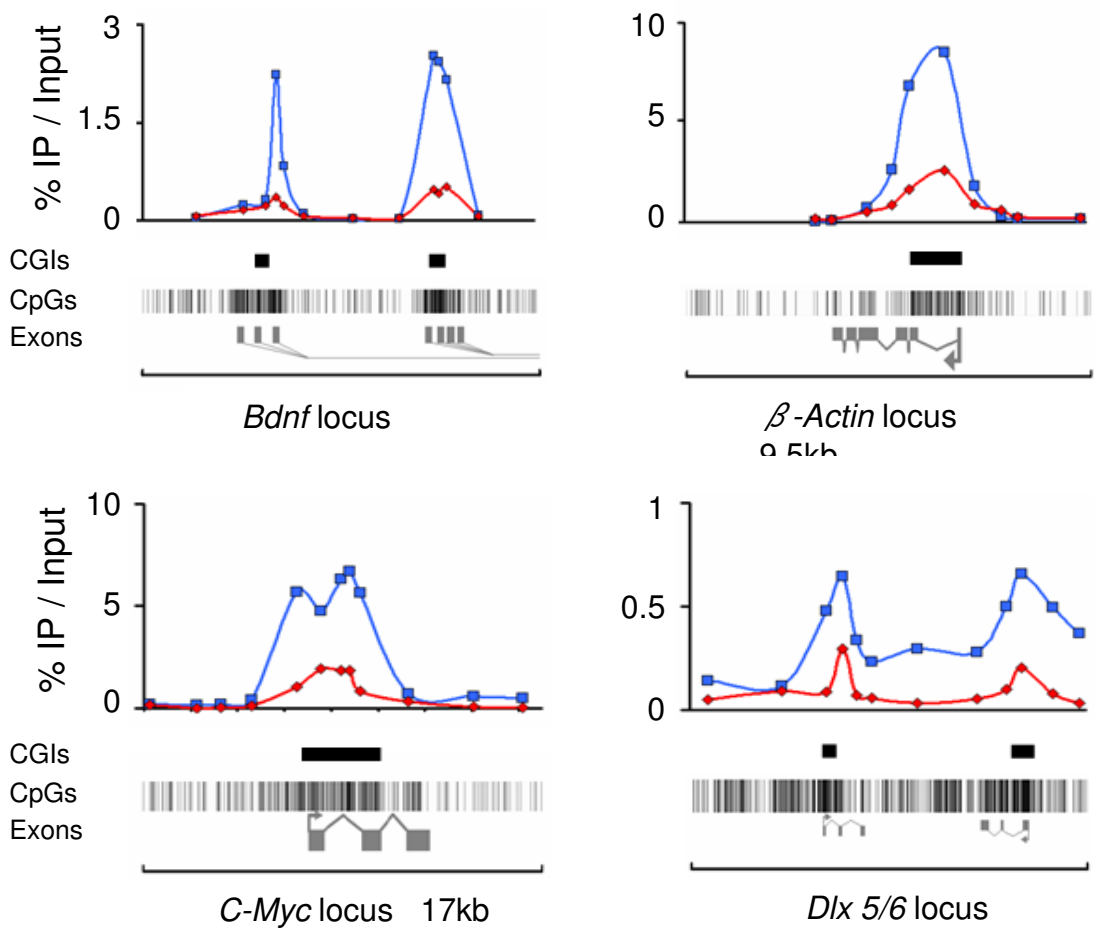

Figure 6.2-2. H3K4me3 modified histone tails are depleted over CGIs in sh*Cfp1* mix3 cell lines. ChIP-PCR profiles for four loci containing 6 CGIs. pSuper10 Control 3T3 cells (blue lines) once again reveal patterns of binding similar to those in the mouse brain. These peaks of H3K4me3 are dramatically reduced in the Cfp1 deficient cell line "mix3" (red lines).

General figure layout and abbreviations follow those outlined in figure 4.2-1.

6.2.3 H3K27me3 and H3K9me3 profiles over CGIs

The histone modifications H3K27me3 and H3K9me3 were also investigated over the CGI loci outlined above. Genome wide analysis reveals that H3K27me3 is enriched over less discrete regions than the H3K4me3 modification. Analysis of the H3K27me3 ChIP-Seq data generated by the Bernstein lab (Mikkelsen, Hanna et al. 2008) reveals that this modification is not frequently found over CGIs (chapter 4.4.4). Notable exceptions to this rule included a minority of islands which lack both Cfp1 and H3K4me3.

Analysis of H3K27me3 patterns over the *Bdnf* promoter locus in pSuper10 control 3T3 cells found no identifiable peaks of enrichment for this histone modification (Figure 6.2-3). Subsequent analysis of the *β-actin, C-Myc* and *Dlx5/6* loci also reveal little or no enrichment of this modification over either the CGI or flanking regions. Cells deficient for Cfp1 did not display any changes in the H3K27me3 profiles to those seen in the vector control cells (Figure 6.2.3, A). As a further control H3K9me3 levels were analysed over the CGIs at the *Bdnf* and *β-actin* loci. In both cases the H3K9me3 profiles revealed no enrichment over the predicted CGIs. These patterns remained low and consistent in both the vector control and the Cfp1 depleted cell lines (Figure 6.2-3, B).

In conclusion, the reduction of H3K4me3 signal in the Cfp1 depleted cells (without a reduction of the H3K27me3 and H3K9me3 signal) furthers the argument for a specific link between the active histone modification and the CXXC protein, possibly facilitated through the Set1a and Set1b complexes of which Cfp1 is a core component.

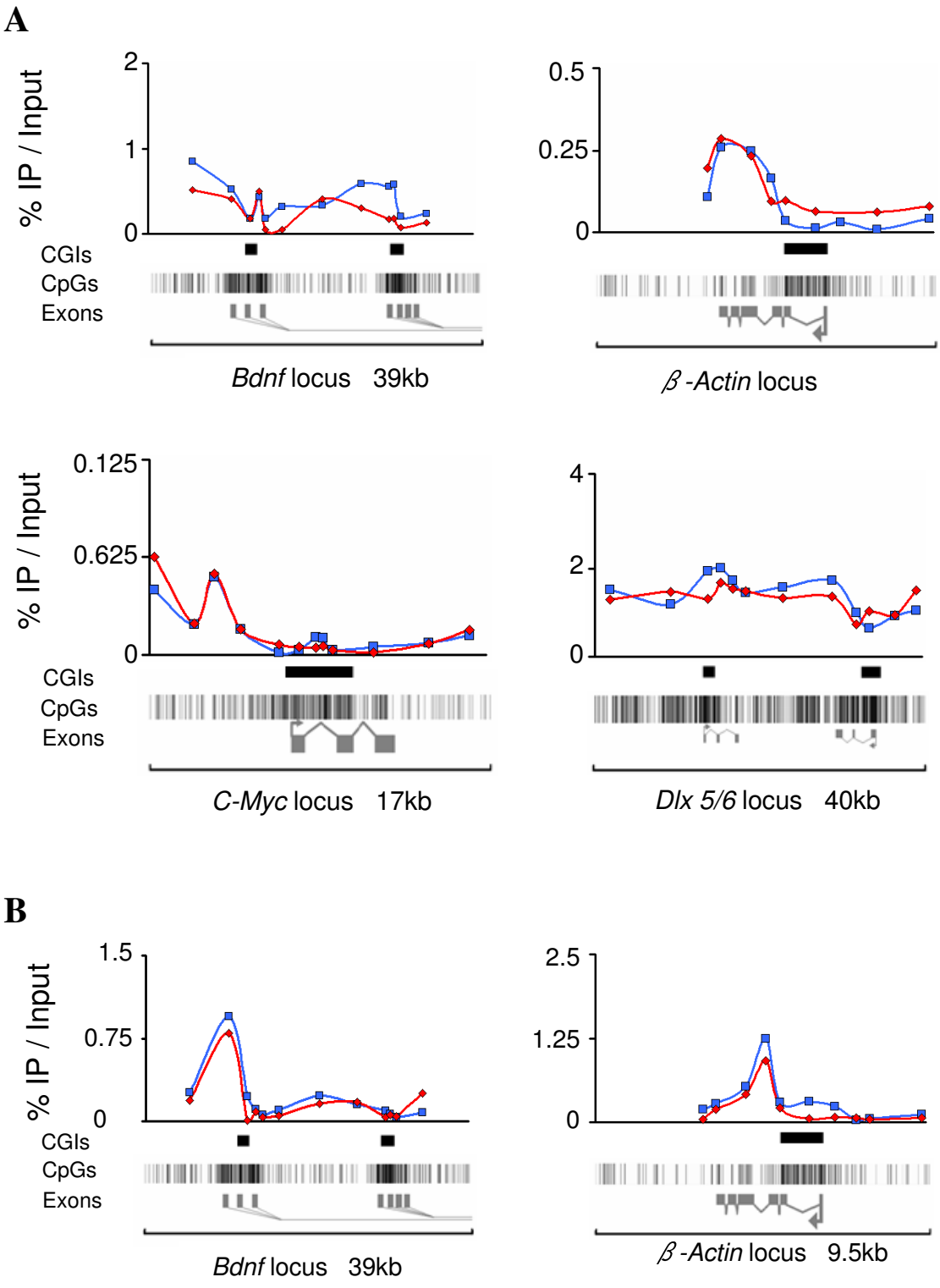

Figure 6.2-3. Patterns of H3K27me3 and H3K9me3 modified histone tails are unaffected in sh*Cfp1* mix3 cell lines. ChIP-PCR profiles for H3K27me3 (6 CGIs, **A**) and H3K9me3 (3 CGIs, **B**). Both H3K27me3 and H3K9me3 patterns in the pSuper10 Control 3T3 cells (blue lines) closely match those seen in the Cfp1 deficient cell line "mix3" (red lines). As such loss of Cfp1 does not affect H3K27me3 levels globally or locally at CGIs.

6.2.4 Comparisons of 2 individual shRNA constructs by ChIP-PCR

In order to ensure that the ChIP profiles seen in the "shMix3" cell line were not due to off target effects, ChIP-PCR was carried out over the *Bdnf* and *β-actin* loci on cell lines transfected with just the sh986, sh1250 or sh1920 constructs. Earlier immunoblotting had revealed that the cell lines 986 clone 4 and 1920 clone 4 both resulted in high levels of Cfp1 depletion (Figure 6.1-2). The sh1250 construct was not able to reduce levels of the Cfp1 protein (Figure 6.1-2) which was verified through the ChIP-PCR profiles (data not shown).

ChIP-PCR profiles for Cfp1, H3K4me3 and H3K27me3 were carried out in both the vector control "pSuper10" cell line as well as in the sh986 cl4 and sh1920 cl4 cell lines (Figure 6.2-4). Both of these cell lines show high levels of reduction in the Cfp1 signal over CGIs when compared to the pSuper10 vector control cells. Interestingly the sh986 cl4 cells revealed lower levels of Cfp1 reduction than the 1920 cl4 cell line. Furthermore, both of these cell lines contain higher levels of Cfp1 than the "shMix3" cell line possibly due to a combinatorial effect of the two short hairpin sequences.

Both the "sh986 cl4" and "sh1920 cl4" cell lines reveal a precipitous drop in the levels of the H3K4me3 modification at CGIs when compared to the vector control cell line. This analysis also reveals that the level of Cfp1 binding is somewhat proportional to the histone H3K4me3 modification over the islands. This is most pronounced at the *β-Actin* locus where the sh1920 cell line shows both increased Cfp1 depletion and H3K4me3 reduction when compared to the sh986 cell line (Figure 6.2.4). Analysis of the H3K27me3 modification over these loci in the two Cfp1 deficient and vector control cell lines reinforced the earlier results in the shMix3 cells. Both the vector control and the individually transfected cell lines failed to contain peaks of the H3K27me3 modification over the regions profiled, once more distancing the H3K27me3 modification from the Cfp1 protein.

In conclusion, the results of the earlier ChIP-PCR experiments of the "shMix3" cells were repeated using two independent short hairpin constructs. As such the knockdown effect is likely due to a joint combination of both of the constructs and not due to off target effects.

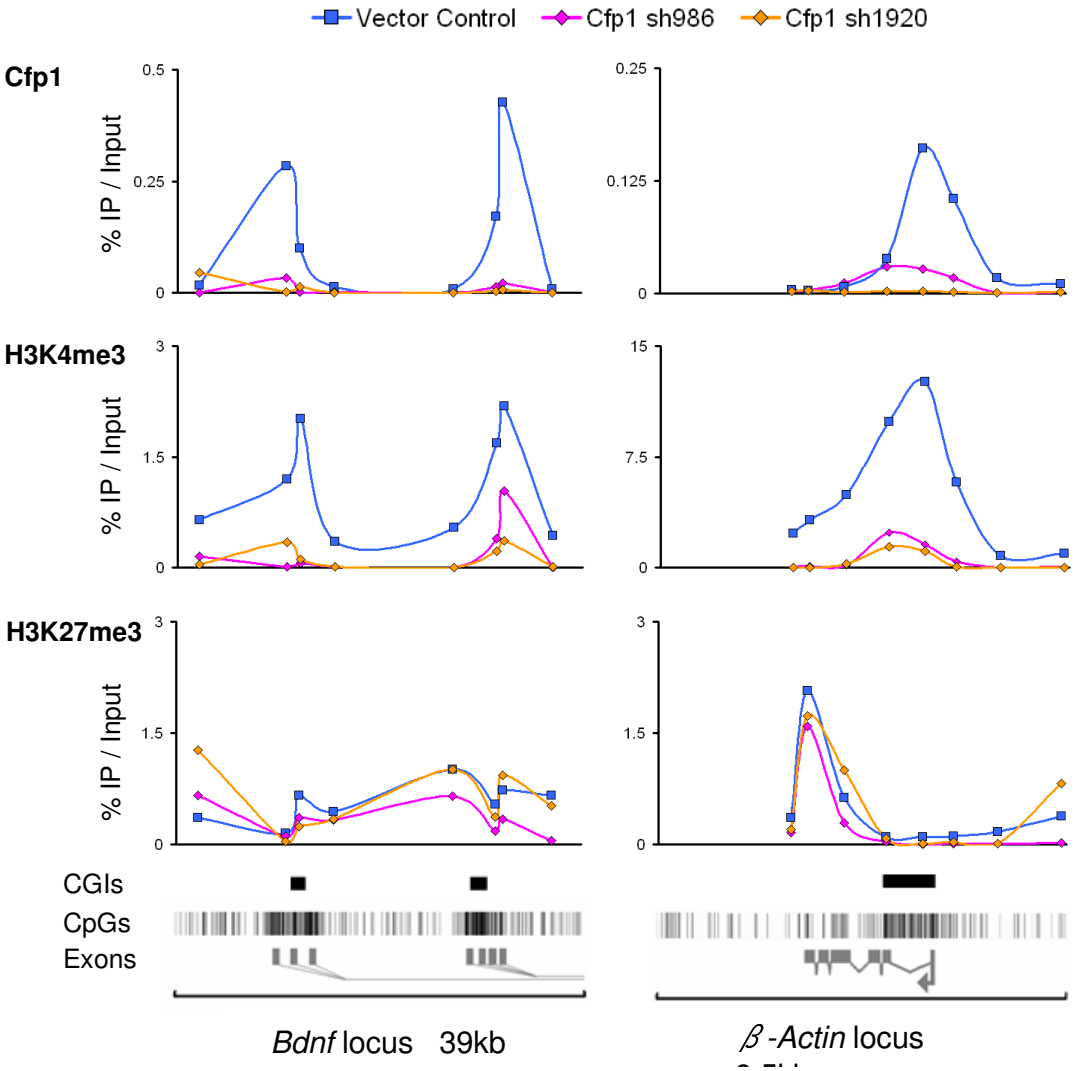

Figure 6.2-4. Two independent Cfp1 deficient cell lines derived from individual shRNAs give identical results to the shMix3 cell line. ChIP-PCR profiles for Cfp1, H3K4me3 and H3K27me3 repeat the earlier findings in which a mixed shRNA cell line was used. pSuper10 cell lines were compared to both 986 clone 4 and 1920 clone 4 cell lines. Each individual sh*Cfp1* cell line shows a reduction in both Cfp1 and H3K4me3 levels comparable to the mixed cell line. Again the H3K27me3 patterns are unchanged at both *bdnf* and *β-actin*. Cfp1 immunoprecipitations were carried using a different antibody from the previous ChIP-PCR studies ("Cfp1-Skalnik" *rabbit polyclonal*). As such these experiments not only verify the individual knockdown cell lines but also the specificity of two separate antibodies for Cfp1.

General figure layout and abbreviations follow those outlined in figure 4.2-1.

6.3 Long term culture of Cfp1 shRNA cells

Although transient transfection is often used to reduce protein levels in the short term, the decision was taken to use short hairpin RNA constructs to stably reduce levels of Cfp1. This should result in a long term reduction in the Cfp1 transcript allowing thorough investigations to be carried without a reduction in silencing. However, under particularly stressful conditions such as loss or reduction of important proteins or complexes, cells have been known to lose this silencing whilst maintaining the selectable resistance markers usually associated with successful knockdowns. This behaviour may be arising as selection prefers the faster growing "normal" cells which are able to escape suppression and become the predominant cell type with respect to the slow growing deficient cells. This was indeed the case during long term culture of the Cfp1 deficient cell lines.

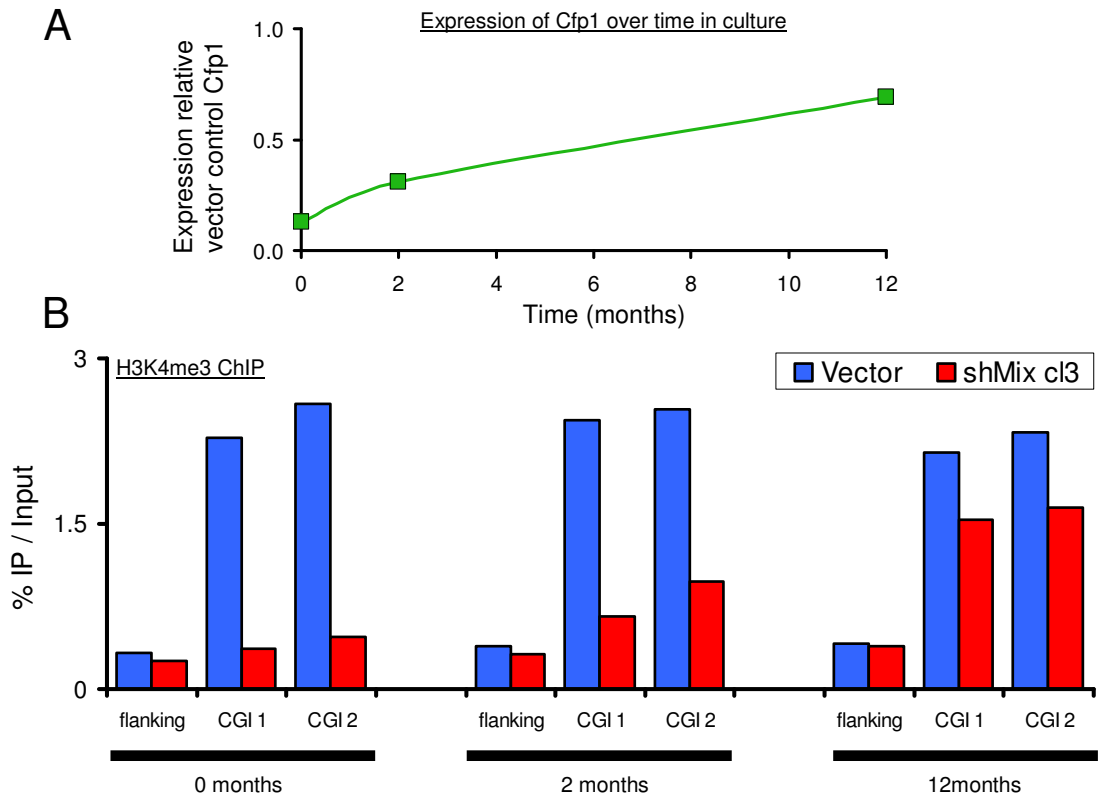

Figure 6.3-1. Reversion of sh*Cfp1* cell lines. A. Expression of Cfp1 returns to wild type over time in culture. The graph shows that the initial 87% reduction returns to a 12% reduction over 12 months. Expression levels of Cfp1 are normalised to Gapdh. **B**. ChIP-PCR for H3K4me3 over the 2 CGIs and a flanking region of the *bdnf* promoter. As Cfp1 levels return to wild type (**A**) so do the H3K4me3 levels over the islands. Vector control cells are represented by blue bars, sh*Cfp1* mix3 cells represented by red bars.

Levels of Cfp1 transcript were analysed through qPCR quantification at certain time points over a twelve month culture period. Levels of Cfp1 expression increase from 13% to 70% of wild type over this time period with reversion occurring as quickly as two months (Figure 6.3-1, A). The increase in Cfp1 was matched by an increase in the H3K4me3 levels seen over the islands, once again highlighting a potential link between this modification and Cfp1. These ChIP results also reveal that the H3K4me3 modifications over CGIs can be restored upon reintroduction of the Cfp1 protein (Figure 6.3-1). As the magnitude of the changes in the Cfp1 and the H3K4me3 signals appear to be closely linked this might represent a direct interaction between the CXXC protein and the modification through the Set1 complex.

This reversion to wild type-like levels of Cfp1 was also noted by Skalnik and colleagues during their work on human PLB-985 myeloid cells transfected with a sh*Cfp1* construct. Initial results revealed that the 70% reduction in Cfp1 expression was reduced to just 15% after twenty passages (Young and Skalnik 2007). The authors argue that this is most likely to be as a result of the loss of silencing from the short hairpin construct leading to an "outgrowth of spontaneously rescued cells". Taken together these findings imply that Cfp1 plays an important role in normal cell viability.

6.4 Cfp1 and links to transcription

One of the more prominent models to explain the maintenance of CGIs in an unmethylated state is that these loci are regions high in transcriptional activity. Indeed CGIs are often found at the 5' ends of genes and frequently (76% of CGIs) associated with the binding of the non-phosphorylated form of RNAP II (Figure 4.4-5). Experiments in ES cells reveal that disruption of the transcription factor binding site for Sp1 over the *aprt* gene resulted in loss of expression and subsequent methylation of the gene in resulting mice. As CGIs are found over the TSS of many genes, disruption of CGI binding factors along with the histone modifications may in some way alter the transcription from underlying genes. Indeed the levels of Dnmt1 expression were also reduced to around 50% of wild type in Cfp1 null ES cells (Carlone, Lee et al. 2005). In order to test whether or not expression levels of genes are altered in the sh*Cfp1* cells, the expression from candidate genes were quantified by qPCR analysis

between vector control and knockdown cells. Two of the individually transfected cell lines (sh986 and sh1960) as well as a cell line transfected with mixture of all three constructs show reduction in gene expression across several candidate genes when compared to the vector control cells (Figure 6.4-1). Interestingly the sh1250 cell line did not show any change in expression when compared to the vector controls and this backs up the finding that these cell lines did not show any depletion in the protein levels of cfp1 (data not shown). As the promoter regions of the genes analysed in Figure 6.4-1 were also investigated in the ChIP-PCR experiments (Figures 6.2-1 to 6.2-3) the results of this transcriptional analysis can be linked to changes in the histone modification profiles. The loss of transcription could be as a result of reduced H3K4me3 modifications over the CGIs at these promoters, tying the process of CGI maintenance with a functional role in transcription. These preliminary findings imply that loss of Cfp1 results in a loss of H3K4me3 from many if not all CGIs resulting in a change in the level of transcription from many CGI containing genes.

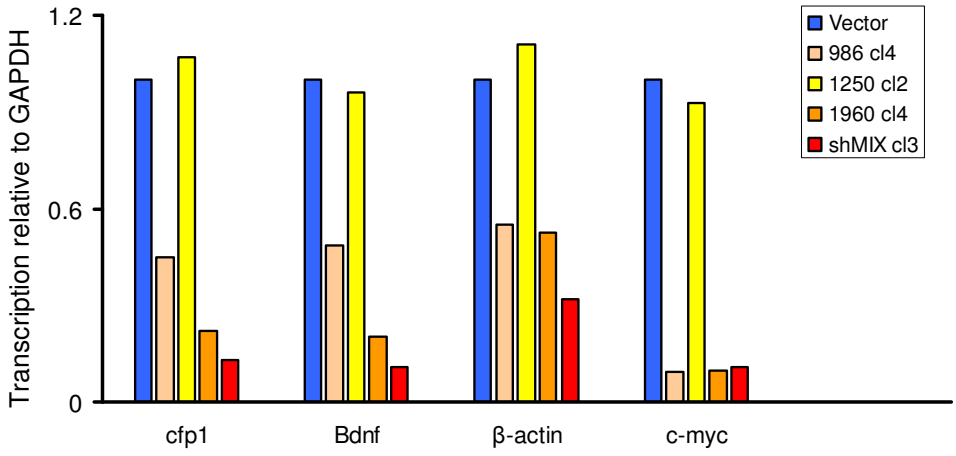

Figure 6.4-1. Gene expression for four genes in sh*Cfp1* and control cell lines. Quantitative PCR results reveal that the expression of several genes is reduced upon reduction of Cfp1. Both the vector control (blue) and sh1250 cl2 (yellow) cell lines contain wild type levels of Cfp1 and shown no discernable differences in the expression of *Cfp1, bdnf, β-actin* or *C-myc*. In contrast, cell lines deficient in Cfp1; 986 cl4 (tan), 1960 cl4 (orange) and shMix cl3 (Lander, Linton et al.) all show reduction in expression at *Cfp1*,*bdnf, β-actin* or *C-myc*.

However, this in itself poses a problem towards the normalisation of the data. As the expression data is normalised to the expression of the housekeeping gene *Gapdh* (which has its own 5' CGI) it is hard to properly quantify the signals. Nevertheless the fact that the s1250 transfected cell line doesn't show any changes in expression whilst the cell lines known to be Cfp1 deficient (sh986, 1920 and mix3) do, argues that transcription is likely reduced in the absence of wild type Cfp1 levels

Cfp1 deficient cell lines generated previously by other laboratories also reveal links between Cfp1 levels to transcription and differentiation. Human myeloid cells stably transfected with a short hairpin construct directed against Cfp1 resulted in a loss of granulocytic differentiation (Young and Skalnik 2007). Restoration of Cfp1 levels also restored the differentiation potential of the cell line. Taken together with the above findings, loss of Cfp1 may be leading to a loss in H3K4me3 levels at promoters, affecting the genes responsible for differentiation.

However, as RNAP II is found at the majority of CGIs alongside Cfp1 and H3K4me3 (see chapter 4.4.3 section iv) one may argue that RNAP II itself or the act of transcription (not Cfp1) is recruiting the H3K4 HMTase complexes such as Set1. To test this further experiment was carried out by Pete Skene and Jim Selfridge in the Bird laboratory in which an artificial promoterless CGI was interrogated for its protein content (Thomson, Skene et al. 2010). The insertion of an enhanced green fluorescent protein (eGFP) reporter gene (containing 60CpGs in 720bp of sequence) and a puromycin resistance gene (containing 93 CpGs in 600bp sequence) into the TβC44 ES line results in the generation of two promoterless and largely unmetylated CGIs (Chambers, Silva et al. 2007; Thomson, Skene et al. 2010). As these CGIs are separated from a promoter they allow for the investigation of the proteins and histone modifications recruited to a CGI in a promoterless environment. As RNAP II is found at the majority of CGIs in our sequencing analysis this experiment can investigate any links between Cfp1 and H3K4me3 at CGIs in the absence of RNAP II (although one may argue that transcription is occurring across the CGI as these sequences are found at the 3' end of the gene for Nanog). Interestingly, the insertion of these two CGI-like sequences was enough to generate two novel H3K4me3 peaks. Furthermore the pattern of H3K4me3 modification followed that of the CXXC protein Cfp1. As this occurs without the recruitment of RNAP II one can argue that Cfp1, and

not RNAP II, is the protein responsible for the recruitment of the H3K4 HMTase complexes such as Set1 to CGIs through specific interactions with its CXXC domain (Thomson, Skene et al. 2010).

Chapter 7: Discussion and future work

The dinucleotide sequence CpG varies in both its modified state and distribution throughout the mammalian genome. Although this dinucleotide is generally found methylated at low density throughout the bulk of the genome, discrete regions rich in non-methylated CpGs form around 1% of the total genomic sequence. Not only are these regions of interest due to their increased CpG densities, they remain stably non-modified in a genome rich in CpG methylation [i]. These regions, termed CpG islands, were found to co-localise to the 5' ends of around 60% of annotated genes (Robert Illingworth, unpublished observations). As the majority of these genes are housekeeping in function this implies that CGIs may have a link with the regulation of a transcriptional state. Indeed CGI methylation has importance in the silencing of genes during X inactivation as well as imprinting (Norris, Brockdorff et al. 1991). Additionally, several genes involved in tumour suppression and cell cycle regulation have been found to acquire methylation of their CGIs (Dobrovic and Simpfendorfer 1997; Stirzaker, Millar et al. 1997; Esteller, Garcia-Foncillas et al. 2000; Virmani, Rathi et al. 2001; Harden, Tokumaru et al. 2003). As CGIs have the ability to become methylated during these processes it suggests that some factor or factors may be protecting these dinucleotides from becoming methylated.

7.1 CGIs may be bound by an array of CGI specific proteins

Previously the study of CGIs has largely focused on the DNA sequence itself and the distribution of these CGIs throughout the genome. However, the majority of work has failed to investigate what proteins, if any, are specifically bound to CGIs. A notable exception to this was a study carried out by Dr Macleod and colleagues in 1994 which investigated the possibility that the transcription factor Sp1 was in fact a CGI binding protein (Macleod, Charlton et al. 1994). In this work it was found that mutation of the Sp1 binding site within the 5' CGI of the *aprt* gene in transgenic mice resulted in a gain of methylation in resulting mice. Subsequent work on the effect of CpG rich DNA sequences to the stability of associated nucleosome raised the possibility that one of the major functions of the CGI

i. Although the majority of CGIs remain methyl-free some are seen to gain methylation in concordinace with gene silencing events (<0.1% genome).

is to destabilise nucleosomes over promoter regions to allow transcription factors (such as Sp1) to bind (Hargreaves, Horng et al. 2009; Ramirez-Carrozzi, Braas et al. 2009; Singh 2009). Although this destabilisation effect may assist in the binding of transcription factors, the work presented in this thesis along with a recent study by Klose and colleagues (Blackledge, Zhou et al. 2010) reveals that CGIs are also bound by a unique set of binding proteins which can in turn affect the local chromatin state. These proteins were identified as potential CGI binding factors due to their highly conserved CXXC zinc finger binding domain which has previously been shown to bind specifically to non-methylated CpG dinucleotides *in vitro* (Voo, Carlone et al. 2000). Further to this, the third CXXC domain from the Mbd1 protein was used previously by Dr Rob Illingworth in the laboratory of Adrian Bird to purify out non-methylated, CpG rich DNA sequences resulting in a high resolution map of the CGIs in both the human and mouse genomes (Illingworth, Kerr et al. 2008; Illingworth, Gruenewald-Schneider et al. 2010).

In mammals around ten proteins contain this CXXC domain (Figure 1.4-1) of which several have roles in the regulation of chromatin state such as the protein DNA methyltransferase Dnmt1 (Pradhan, Esteve et al. 2008), the H3K4MTases Mll1 and Mll2 (Birke, Schreiner et al. 2002) and the H3K36 demethylase Kdm2a (Tsukada, Fang et al. 2006). The work outlined here identifies the CXXC domain containing protein as an important CGI specific binding factor within the mammalian genome. As Cfp1 is a core component of the Set1 H3K4HMTase, this protein appears to link the underlying DNA sequence rich in CpG dinucleotides, to changes in the local chromatin state through modification of histone tails.

Early work by Dr Skalnik revealed that Cfp1 can bind to non-methylated CpG rich probes *in vitro* (Voo, Carlone et al. 2000) which was verified in this thesis by *in vivo* ChIP-PCR experiments over the *Xist* locus. Although other laboratories have studied the Cfp1 protein none had proposed that it may be able to bind to CGIs. This was somewhat surprising given the fact that its CXXC domain was known to have a preference for non-methylated CpG rich sequences. Instead the majority of work on Cfp1 has focused on its interactions with other proteins such as the Set1a (Lee and Skalnik 2005) and Set1b (Lee, Tate et al. 2007) complexes or the Dnmt1 protein (Butler, Lee et al. 2008).

The finding that both Cfp1 and Kdm2a bind to CGIs raises the possibility that the other CXXC domain containing proteins may also behave similarly. Indeed Mll1 was shown to bind over CGIs and has previously been shown to bind preferentially to non-methylated CpG rich DNA by its CXXC domain (Birke, Schreiner et al. 2002). It will be interesting to find out whether Tet1, a protein with demethylase activity as well as a CXXC domain, can bind to CGIs in the same way. These CXXC domain containing proteins appear to be acting in a similar yet opposing manner to the methyl binding domain (MBD) containing proteins; each binding to a particular state of CpG within the genome. This ability to bind and interpret modified or non-modified forms of the cytosine base may have arisen during evolution as an efficient method to increase the coding potential of the mamalian genome.

7.2 The effects of Cfp1 binding at CGIs

The evidence presented in this thesis suggests a significant association between the Cfp1 protein and the H3K4me3 modification at CGIs. Although the high resolution profiles for several histone methylations such as H3K4me3 and H3K27me3 have been collected (Barski, Cuddapah et al. 2007; Mikkelsen, Ku et al. 2007) this analysis has not been restricted to CGIs themselves. Upon doing so it appears that in addition to containing a unique set of proteins such as Cfp1 and Kdm2a, CGIs are also unique in the histone modifications present. This was most noticeable for the enrichment of H3K4me3 tails and depletion of H3K9me3 and H3K27me3 tails directly over CGIs.

One of the most striking findings of this work has been the intimate association between the CXXC domain containing protein Cfp1 and the active histone modification H3K4me3. Both marks were found throughout the genome to co-localise to CGIs and depletion of the Cfp1 protein was followed closely by levels of H3K4me3 modifications over CGIs. This direct link appears to be facilitated through the enzymatic complex responsible for setting up the levels of the histone H3K4me3 modification. Cfp1 is in fact a subunit of the Set1a/b and yeast COMPASS H3K4 HMTase complexes responsible for the setting up and maintenance of the H3K4me3 modification (Lee and Skalnik 2005; Lee, Tate et al. 2007; Takahashi, Lee et al. 2009). The homologue of Cfp1 in yeast (*Spp1*) is not

essential for the overall viability of the yeast Set1 complex (Takahashi, Lee et al. 2009); however removal of *Spp1* results in loss of H3K4me3 levels.

Interestingly *Spp1* in yeast lacks the CXXC domain found in its mammalian counterpart, possibly as this organism lacks DNA methylation. Therefore the CXXC domain may have arisen in organisms containing DNA methylation as a means of targeting the Set1 complex to specific loci (non-methylated CpG rich regions often occurring at the 5' ends of genes: CGIs). Indeed sequence alignments of the Cfp1 CXXC domain from several organisms reveals high levels of conservation within the organisms in which DNA methylation has been detected (Figure 5.2-2). In this regards, the CXXC domain within Cfp1 may target the Set1 HMTase complex to CGIs resulting in H3K4me3 at these loci (Figure 7.2-1). Loss of this targeting to the 1% of the genome occupied by CGIs would reduce the occupancy of the complex at such sites resulting in a CGI specific loss of H3K4me3 marks without a disruption on a global level.

The functional significance of the H3K4me3 modification at CGIs is largely unknown however it is largely thought to be required to form transcriptionally permissive chromatin (Bannister, Schneider et al. 2002; Bernstein, Humphrey et al. 2002; Santos-Rosa, Schneider et al. 2002; Schubeler, MacAlpine et al. 2004). As around 60% of promoters contain a CGI, these regions (and their associated H3K4me3 modifications) may be important in the regulation of transcription. As the promoters of housekeeping genes contain an associated CGI, these non methylated CpG rich regions may be required to ensure high levels of transcription.. Although poorly understood, it has been proposed that the methylation of H3K4 residues may facilitate the formation of a euchromatic state through the attraction of specific H3K4me3 binding factors (Ruthenburg, Allis et al. 2007). Several of these H3K4me3 binding proteins have roles in epigenetic regulation such as the protein ING which attracts histone acetyltransferase complexes or BPTF which is part of the NURF nucleosomal remodelling complex (Mizuguchi, Tsukiyama et al. 1997) and TFIID which is a member of the transcriptional machinery associated with RNAPII (Vermeulen, Mulder et al. 2007). The latter of these H3K4me3 binding proteins may represent a link to the recruitment of the RNAPII complex at CGIs (Figure 7.2-1, B). The enrichment of H3K4me3 modified histone tails over CGIs may also act to reinforce the

binding of Cfp1/Set1 complex at these loci. Studies in yeast have found that the homologue of Cfp1, *Spp1*, contains a PHD domain with a particular preference to histones H3 tails containing the me2 and me3 modifications (Murton, Chin et al. 2010)

Alternatively, H3K4me3 modified tails may be refractory to the binding of certain proteins or occupancy by "silencing" histone modifications. Indeed the *de novo* DNA methyltransferase cofactor Dnmt3L is unable to bind to histone tails with this modification *in vitro* (Ooi, Qiu et al. 2007). As such, regions of the genome rich in the H3K4me3 modification such as CGIs would remain inaccessible to these methyltransferases and would thus remain free of methylation (Figure 7.2-1, A). This may therefore go some way to explain how CGIs remain methylation free in a genome which is largely methylated. Further analysis of the methylation status of the CGIs in the Cfp1 deficient cell line should result in a greater understanding of this model.

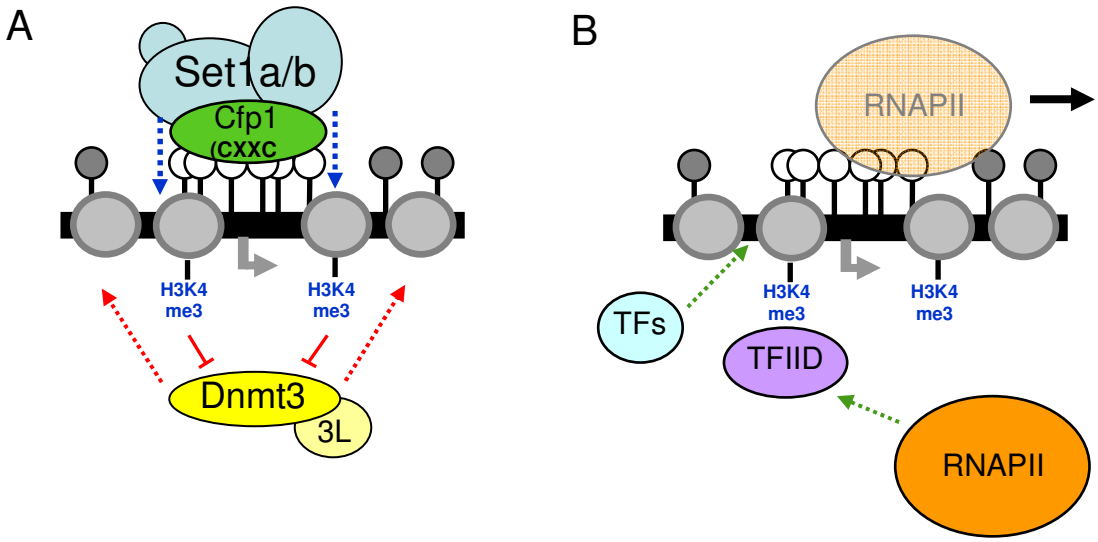

Figure 7.2-1. Potential models for the maintenance of CGI chromatin state. Figure **A.** shows a proposed model whereby the underlying CpG rich non-methylated DNA influences the overlying chromatin state. Cfp1 (green) is proposed to bind to non methylated CGIs through interactions with its CXXC domain. As Cfp1 is a part of the Set1a/b complexes (light blue) this results in the trimethylation of H3K4 residues on the surrounding histones (histone octamers represented by grey circles). As histone tails containing this modification has been shown to be antagonistic to the binding of Dnmt3 and 3L (yellow) this may represent a model for the maintenance of a non-methylated CGI state. Figure **B**. represents how the H3K4me3 modification at CGIs may affect transcription and the recruitment of RNAPII. Specific proteins such as the TFIID factor (purple) bind to H3K4me3 modified tails and may in conjunction with the binding of specific transcription factors ("TFs – light blue") to recruit RNAPII to islands. Therefore both the non methylated CpG dinucleotide and the H3K4me3 modified histone tails act as platforms for the binding of chromatin modifying proteins and the transcriptional machinery.

As CGIs are often found associated with the 5' ends of genes it was not surprising to find that many of these CGIs co-localised to a peak of RNAP II. However this co-localisation with peaks of H3K4me3 at CGIs may reflect indirect recruitment of the Set1 complexes by RNAPII itself. Another of the mammalian H3K4 HMTases, MLL1 (which also contains a CXXC domain) was found to be tightly associated with the RNAP II protein at the 5' ends of active genes (Guenther, Jenner et al. 2005). The ChIP-Sequencing data presented above revealed that around 7% of the Cfp1/H3K4me3 rich CGIs were seen to lack RNAPII. This argues that the transcriptional machinery is not a required element for the generation of the H3K4me3 peaks over islands. This was tested further by Pete Skene and Jim Selfridge in the Bird Laboratory by investigating promoterless "CGI-like" constructs in ES cells. Using the TβC44 ES cell line (Chambers, Silva et al. 2007) in which the genes encoding eGFP (containing 60 CpGs in a 720bp sequence) or Puromycin (93 CpGs in a 600bp sequence) had been inserted into the genome resulted in DNA sequences with the typical CpG density of a CGI with no associated promoter. ChIP-PCR over these promoterless CGIs revealed that the non methylated CpG rich sequence was sufficient to recruit Cfp1 and to generate a peak of H3K4me3 in the absence of RNAPII (Thomson, Skene et al. 2010). As proposed in Figure 7.2-1 B, RNAPII may be recruited to CGIs through association with the H3K4me3 binding factor TFIID as well as specific transcription factors. This model would go some way to describing the finding that 88% of Cfp1/H3K4me3 enriched islands also contain the RNAPII protein and suggests that CGIs may have important roles in the preparation of promoter regions (at least for the housekeeping genes) for transcription.

7.3 The ES cell discrepancy

The bulk of the investigations described above were carried out in somatic cells of the mouse brain or cultured NIH-3T3 fibroblasts. However, to better understand how these CGIs are initially created it may be important to study Cfp1 in the context of the pluripotent stem cell. Cfp1 -/- ES cells have been generated in the lab of Dr Skalnik which reportedly contain a four fold increase in the levels of H3K4me3 when compared to wild type cells. This is somewhat confusing as our findings along with those of Ansari and colleagues reveal a loss of H3K4me3 from CGI loci. This may be in part due to

the difference in Cfp1 depletion as the work outlined above and studies carried out by Ansari and colleagues reduced Cfp1 levels by antisense transcription whereas Skalnik and colleagues generated a knock out cell line. However, this may also be due to the fact that Skalnik's work was carried out on ES cells compared to the HEK293 cells used by Ansari and colleagues and the somatic cells used in our own work.

To test if there were any differences in the behaviour of Cfp1 in ES cells, a series of experiments were carried out by Thomas Clouaire in the Bird laboratory. ChIP-PCR in WT ES cells gave results similar to those shown above for the 3T3 cells (with peaks of binding occurring over the CGIs), however when repeated on Cfp1 null ES cells (gifted by Dr Skalnik) this signal failed to disappear. As these immunoprecipitations were carried out with the same antibody which detected the reduction of Cfp1 levels in the sh*Cfp1* 3T3 cells this was a surprising result raising the question as to what the antibody was recognising in ES cells. To ensure that this finding was an ES cell specific issue, the earlier ChIP-PCR studies on the sh*Cfp1*cells were repeated using an antibody generated by Dr Skalnik raised against the full length protein. This antibody gave similar results to those seen using the commercially available "H-120" antibody with peaks of enrichment for Cfp1 over the CGIs lost in the knockdown cell line (Figure 6.2-4). Genome wide ChIP-Seq with the Skalnik antibody generated data sets which closely matched that of the initial antibody (Figure 4.4-3). Therefore it can be concluded that for somatic cells both the H-120 commercial antibody and the full length Skalnik antibody specifically recognise Cfp1. However, as studies in the Cfp1 null ES cells using the Skalnik antibody resulted in the expected loss of signal over CGIs, this raises the possibility that the H-120 antibody is non specific in ES cells. As this antibody is raised to a portion of the CXXC domain it might represent the detection of another CXXC protein specifically in ES cells. Further work into any ES cell specific CGI factors will go some way to better understanding how these islands are initially set up.

7.5 Concluding remarks

The dinucleotide sequence CpG can be found in both modified and non-modified forms throughout the mammalian genome, the distribution of which increases the coding potential of the genome. Typically CpG is found to contain a methylated cytosine base at low density (typically 1 per 100bp)

throughout the majority (~99%) of the genome. This modified base is bound by a group of proteins containing methyl binding domains which in turn recruit enzymatic complexes capable of regulating the local chromatin state. Typically these MBD recruited proteins aid the formation of a repressive heterochromatic state, linking the modified CpG dinucleotide to a compact chromatin environment. In contrast to these methylated CpGs, a fraction of the genomic sequence (approximately 1-2% of the total sequence) is seen to contain high densities of the non methylated form of the CpG dinucleotide over approximately 1kb stretches of the genome (known as the CpG islands). Furthermore, these CGIs are often found associated with the 5' ends of annotated protein coding genes. In the same way that the methylated CpG dinucleotide recruits MBD containing proteins, the work presented here reveals a group of proteins containing a CXXC domain which bind specifically to the non-methylated CpG dinucleotide. Several of these CXXC proteins have roles in the modulation of the chromatin environment. The protein Cfp1 for example is a component of the Set1a and Set1b H3K4 HMTase complex responsible for the formation of euchromatin. Additionally the CXXC proteins Mll1 and Kdm2a are important for the modulation of the H3K4me3 and a H3K36me3 marks respectively. As such these non-methylated islands may be acting as "beacons" for the recruitment and binding of factors which modify the chromatin state, simplifying the genome by the restriction of potential binding sites to around 1% of the total sequence.

The majority of the work presented here focused upon the protein Cfp1. Although this protein was shown to localise to CGIs and be sufficient to establish chromatin rich in the H3K4me3 modification the precise mode by which this is accomplished is still unknown. The hypothesis raised is that the Set1a and Set1b complexes are recruited to CGIs by Cfp1, resulting in the targeting of the complex specifically to non methylated CpGs. Indeed comparisons between Cfp1 and its yeast homologue, *Spp1* reveal a similar function within an H3K4 HMTase complex. As the yeast lacks methylation, the yeast protein lacks any CXXC motif for potential targeting to non-methylated "beacons". One could argue that in the shRNA experiments (chapter 6) depletion of Cfp1 is not resulting in a loss of targeting of Set1 complexes but in a reduced functionality of stability of the complex itself. This would explain the loss of the H3K4me3 peaks over the CGS however it would not account for the near wild type levels of global H3K4me3 seen (Figure 6.1-4). To test this possibility, the CGIs of cell

lines containing Cfp1 DNA binding mutants (with mutations in the CXXC domain to remove any CXXC specific targeting to islands) could be investigated. If H3K4me3 modifications are lost from the islands in the DNA binding mutant cell lines it would indicate a CXXC targeting model for the Set1a/b complex. Further experiments could be devised in which the sh*Cfp1* cells could be investigated for potential rescue of Cfp1. By adding back either wild type or a DNA binding mutant of Cfp1 one could investigate the effects on H3K4me3 levels at islands. If for example the introduction of a DNA binding mutant of Cfp1 was able to restore H3K4me3 levels at islands it would argue that the loss of Cfp1 targeting is not the main reason for the reduction in H3K4me3 peaks seen in the sh*Cfp1* cell lines. Investigations into the precise roles for Cfp1 within the Set1a/b complex are vital in the understanding of the functional importance of the proteins associated with CpG islands.

In addition to these studies it is important to attempt to understand the roles of additional factors at CGIs such as the possible involvement of another CXXC protein, Mll1. As this H3K4me3 HMTase complex was also seen to peak to CGIs, although mainly over promoter regions, it may also be responsible for the maintenance of the overlying epigenetic state. Indeed work by Ansari and colleagues found that shRNA depletion of Cfp1 also reduced the binding of the Mll1 protein at the Hoxa7 CGI along with a reduction of the H3K4me3 signal (Ansari, Mishra et al. 2008).

Overall the findings of this work reveal that the CXXC protein Cfp1 is found specifically at CpG islands. Through its specific non methylated DNA binding ability this protein is recruited to the CpG islands, modulating histone modifications at such sites. As such this represents a mode whereby DNA sequence can drive the overlying chromatin state. Further investigations of Cfp1 and its role at CpG islands should lead to a greater understanding of DNA driven epigenetic control at these loci.

References

Agius, F., A. Kapoor, et al. (2006). "Role of the Arabidopsis DNA glycosylase/lyase ROS1 in active DNA demethylation." Proc Natl Acad Sci U S A **103**(31): 11796-801.

Allen, M. D., C. G. Grummitt, et al. (2006). "Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase." EMBO J **25**(19): 4503-12.

Ansari, K. I., B. P. Mishra, et al. (2008). "Human CpG binding protein interacts with MLL1, MLL2 and hSet1 and regulates Hox gene expression." Biochim Biophys Acta **1779**(1): 66-73.

Antequera, F. and A. Bird (1993). "Number of CpG islands and genes in human and mouse." Proc Natl Acad Sci U S A **90**(24): 11995-9.

Antequera, F. and A. Bird (1999). "CpG islands as genomic footprints of promoters that are associated with replication origins." Curr Biol **9**(17): R661-7.

Antequera, F., D. Macleod, et al. (1989). "Specific protection of methylated CpGs in mammalian nuclei." Cell **58**(3): 509-17.

Awad, S. and A. H. Hassan (2008). "The Swi2/Snf2 bromodomain is important for the full binding and remodeling activity of the SWI/SNF complex on H3- and H4-acetylated nucleosomes." Ann N Y Acad Sci **1138**: 366-75.

Ayton, P. M. and M. L. Cleary (2001). "Molecular mechanisms of leukemogenesis mediated by MLL fusion proteins." Oncogene **20**(40): 5695-707.

Bannister, A. J., R. Schneider, et al. (2002). "Histone methylation: dynamic or static?" Cell **109**(7): 801-6.

Bannister, A. J., P. Zegerman, et al. (2001). "Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain." Nature **410**(6824): 120-4.

Barreto, G., A. Schafer, et al. (2007). "Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation." Nature **445**(7128): 671-5.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-37.

Bender, J. (2004). "DNA methylation of the endogenous PAI genes in Arabidopsis." Cold Spring Harb Symp Quant Biol **69**: 145-53.

Bernstein, B. E., E. L. Humphrey, et al. (2002). "Methylation of histone H3 Lys 4 in coding regions of active genes." Proc Natl Acad Sci U S A **99**(13): 8695-700.

Bernstein, B. E., M. Kamal, et al. (2005). "Genomic maps and comparative analysis of histone modifications in human and mouse." Cell **120**(2): 169-81.

Bernstein, B. E., T. S. Mikkelsen, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell **125**(2): 315-26.

Bestor, T. H. (1990). "DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes." Philos Trans R Soc Lond B Biol Sci **326**(1235): 179-87.

Bestor, T. H. (1992). "Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain." EMBO J **11**(7): 2611-7.

Bestor, T. H. (2000). "The DNA methyltransferases of mammals." Hum Mol Genet **9**(16): 2395-402.

Bestor, T. H. and V. M. Ingram (1983). "Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA." Proc Natl Acad Sci U S A **80**(18): 5559-63.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.

Bird, A., M. Taggart, et al. (1985). "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA." Cell **40**(1): 91-9.

Bird, A. P. (1978). "Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern." J Mol Biol **118**(1): 49-60.

Bird, A. P. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Res **8**(7): 1499-504.

Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature **321**(6067): 209-13.

Bird, A. P. (1996). "The relationship of DNA methylation to cancer." Cancer Surv **28**: 87-101.

Bird, A. P. and A. P. Wolffe (1999). "Methylation-induced repression--belts, braces, and chromatin." Cell **99**(5): 451-4.

Birke, M., S. Schreiner, et al. (2002). "The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation." Nucleic Acids Res **30**(4): 958-65.

Birve, A., A. K. Sengupta, et al. (2001). "Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants." Development **128**(17): 3371-9.

Blackledge, N. P., J. C. Zhou, et al. (2010). "CpG islands recruit a histone H3 lysine 36 demethylase." Mol Cell **38**(2): 179-90.

Bourc'his, D., G. L. Xu, et al. (2001). "Dnmt3L and the establishment of maternal genomic imprints." Science **294**(5551): 2536-9.

Boyer, L. A., K. Plath, et al. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells." Nature **441**(7091): 349-53.

Braghetti, A., G. Piazzi, et al. (1993). "Multiple DNA-protein interactions at the CpG island of the human pseudoautosomal gene MIC2." Somat Cell Mol Genet **19**(1): 51-63.

Butler, J. S., J. H. Lee, et al. (2008). "CFP1 interacts with DNMT1 independently of association with the Setd1 Histone H3K4 methyltransferase complexes." DNA Cell Biol **27**(10): 533-43.

Byrd, K. N. and A. Shearn (2003). "ASH1, a Drosophila trithorax group protein, is required for methylation of lysine 4 residues on histone H3." Proc Natl Acad Sci U S A **100**(20): 11535-40.

Cao, R., L. Wang, et al. (2002). "Role of histone H3 lysine 27 methylation in Polycomb-group silencing." Science **298**(5595): 1039-43.

Carlone, D. L., J. H. Lee, et al. (2005). "Reduced genomic cytosine methylation and defective cellular differentiation in embryonic stem cells lacking CpG binding protein." Mol Cell Biol **25**(12): 4881-91.

Carlone, D. L. and D. G. Skalnik (2001). "CpG binding protein is crucial for early embryonic development." Mol Cell Biol **21**(22): 7601-6.

Cedar, H., L. Lande-Diner, et al. (2007). "Role of DNA methylation in stable gene repression." J Biol Chem **282**(16): 12194-200.

Chambers, I., J. Silva, et al. (2007). "Nanog safeguards pluripotency and mediates germline development." Nature **450**(7173): 1230-4.

Chuang, L. S., H. I. Ian, et al. (1997). "Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1." Science **277**(5334): 1996-2000.

Cooper, D. N., M. H. Taggart, et al. (1983). "Unmethylated domains in vertebrate DNA." Nucleic Acids Res **11**(3): 647-58.

Cronn, R., A. Liston, et al. (2008). "Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology." Nucleic Acids Res **36**(19): e122.

Dignam, J. D., R. M. Lebovitz, et al. (1983). "Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei." Nucleic Acids Res **11**(5): 1475-89.

Dobrovic, A. and D. Simpfendorfer (1997). "Methylation of the BRCA1 gene in sporadic breast cancer." Cancer Res **57**(16): 3347-50.

Dong, A., J. A. Yoder, et al. (2001). "Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA." Nucleic Acids Res **29**(2): 439-48.

Dou, Y., T. A. Milne, et al. (2006). "Regulation of MLL1 H3K4 methyltransferase activity by its core components." Nat Struct Mol Biol **13**(8): 713-9.

Eckhardt, F., J. Lewin, et al. (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." Nat Genet **38**(12): 1378-85.

Egger, G., G. Liang, et al. (2004). "Epigenetics in human disease and prospects for epigenetic therapy." Nature **429**(6990): 457-63.

Ernst, P., J. K. Fisher, et al. (2004). "Definitive hematopoiesis requires the mixed-lineage leukemia gene." Dev Cell **6**(3): 437-43.

Esteller, M., J. Garcia-Foncillas, et al. (2000). "Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents." N Engl J Med **343**(19): 1350-4.

Fatemi, M., A. Hermann, et al. (2001). "The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain

with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA." J Mol Biol **309**(5): 1189-99.

Finnegan, E. J., R. K. Genger, et al. (1998). "DNA Methylation in Plants." Annu Rev Plant Physiol Plant Mol Biol **49**: 223-247.

Gil, J. and G. Peters (2006). "Regulation of the INK4b-ARF-INK4a tumour suppressor locus: all for one or one for all." Nat Rev Mol Cell Biol **7**(9): 667-77.

Glaser, S., J. Schaft, et al. (2006). "Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development." Development **133**(8): 1423-32.

Glozak, M. A., N. Sengupta, et al. (2005). "Acetylation and deacetylation of non-histone proteins." Gene **363**: 15-23.

Goll, M. G. and T. H. Bestor (2005). "Eukaryotic cytosine methyltransferases." Annu Rev Biochem **74**: 481-514.

Guenther, M. G., R. G. Jenner, et al. (2005). "Global and Hox-specific roles for the MLL1 methyltransferase." Proc Natl Acad Sci U S A **102**(24): 8603-8.

Guenther, M. G., S. S. Levine, et al. (2007). "A chromatin landmark and transcription initiation at most promoters in human cells." Cell **130**(1): 77-88.

Harden, S. V., Y. Tokumaru, et al. (2003). "Gene promoter hypermethylation in tumors and lymph nodes of stage I lung cancer patients." Clin Cancer Res **9**(4): 1370-5.

Hargreaves, D. C., T. Horng, et al. (2009). "Control of inducible gene expression by signal-dependent transcriptional elongation." Cell **138**(1): 129-45.

Hata, K., M. Okano, et al. (2002). "Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice." Development **129**(8): 1983-93.

Hebbes, T. R., A. L. Clayton, et al. (1994). "Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain." EMBO J **13**(8): 1823-30.

Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet **39**(3): 311-8.

Hendrich, B. and A. Bird (1998). "Identification and characterization of a family of mammalian methyl-CpG binding proteins." Mol Cell Biol **18**(11): 6538-47.

Hendrich, B., J. Guy, et al. (2001). "Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development." Genes Dev **15**(6): 710-23.

Honda, B. M., G. H. Dixon, et al. (1975). "Sites of in vivo histone methylation in developing trout testis." J Biol Chem **250**(22): 8681-5.

Horsthemke, B. (2007). "Heritable germline epimutations in humans." Nat Genet **39**(5): 573-4; author reply 575-6.

Hutchins, A. S., A. C. Mullen, et al. (2002). "Gene silencing quantitatively controls the function of a developmental trans-activator." Mol Cell **10**(1): 81-91.

Illingworth, R., A. Kerr, et al. (2008). "A novel CpG island set identifies tissue-specific methylation at developmental gene loci." PLoS Biol **6**(1): e22.

Illingworth, R. S., U. Gruenewald-Schneider, et al. (2010). "Orphan CpG islands identify numerous conserved promoters in the mammalian genome." PLoS Genet **6**(9).

Ingvarsdottir, K., C. Edwards, et al. (2007). "Histone H3 K4 demethylation during activation and attenuation of GAL1 transcription in Saccharomyces cerevisiae." Mol Cell Biol **27**(22): 7856-64.

Jahner, D., H. Stuhlmann, et al. (1982). "De novo methylation and expression of retroviral genomes during mouse embryogenesis." Nature **298**(5875): 623-8.

Jin, S. G., C. Guo, et al. (2008). "GADD45A does not promote DNA demethylation." PLoS Genet **4**(3): e1000013.

Jones, P. A. (2002). "DNA methylation and cancer." Oncogene **21**(35): 5358-60.

Jorgensen, H. F., I. Ben-Porath, et al. (2004). "Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains." Mol Cell Biol **24**(8): 3387-95.

Joshi, A. A. and K. Struhl (2005). "Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation." Mol Cell **20**(6): 971-8.

Kan, J., L. Zou, et al. (2008). "Origin recognition complex (ORC) mediates histone 3 lysine 4 methylation through cooperation with Spp1 in Saccharomyces cerevisiae." J Biol Chem **283**(49): 33803-7.

Kangaspeska, S., B. Stride, et al. (2008). "Transient cyclical methylation of promoter DNA." Nature **452**(7183): 112-5.

Kim, T. H., L. O. Barrera, et al. (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-80.

Klose, R. J., E. M. Kallin, et al. (2006). "JmjC-domain-containing proteins and histone demethylation." Nat Rev Genet **7**(9): 715-27.

Kornberg, R. D. (1974). "Chromatin structure: a repeating unit of histones and DNA." Science **184**(139): 868-71.

Kriaucionis, S. and N. Heintz (2009). "The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain." Science **324**(5929): 929-30.

Kurdistani, S. K. and M. Grunstein (2003). "Histone acetylation and deacetylation in yeast." Nat Rev Mol Cell Biol **4**(4): 276-84.

Kurdistani, S. K., S. Tavazoie, et al. (2004). "Mapping global histone acetylation patterns to gene expression." Cell **117**(6): 721-33.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lee, J. H. and D. G. Skalnik (2002). "CpG-binding protein is a nuclear matrix- and euchromatin-associated protein localized to nuclear speckles containing human trithorax. Identification of nuclear matrix targeting signals." J Biol Chem **277**(44): 42259-67.

Lee, J. H. and D. G. Skalnik (2005). "CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex." J Biol Chem **280**(50): 41725-31.

Lee, J. H., C. M. Tate, et al. (2007). "Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex." J Biol Chem **282**(18): 13419-28.

Lee, J. H., K. S. Voo, et al. (2001). "Identification and characterization of the DNA binding domain of CpG-binding protein." J Biol Chem **276**(48): 44669-76.

Lee, J. S. and A. Shilatifard (2007). "A site to remember: H3K36 methylation a mark for histone deacetylation." Mutat Res **618**(1-2): 130-4.

Lei, H., S. P. Oh, et al. (1996). "De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells." Development **122**(10): 3195-205.

Lewis, E. B. (1978). "A gene complex controlling segmentation in Drosophila." Nature **276**(5688): 565-70.

Li, D. and R. Roberts (2001). "WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases." Cell Mol Life Sci **58**(14): 2085-97.

Li, E. (2002). "Chromatin modification and epigenetic reprogramming in mammalian development." Nat Rev Genet **3**(9): 662-73.

Li, E., T. H. Bestor, et al. (1992). "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." Cell **69**(6): 915-26.

Liang, G., J. C. Lin, et al. (2004). "Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome." Proc Natl Acad Sci U S A **101**(19): 7357-62.

Lin, I. G., T. J. Tomzynski, et al. (2000). "Modulation of DNA binding protein affinity directly affects target site demethylation." Mol Cell Biol **20**(7): 2343-9.

Liu, Y., R. Prasad, et al. (2007). "Coordination of steps in single-nucleotide base excision repair mediated by apurinic/apyrimidinic endonuclease 1 and DNA polymerase beta." J Biol Chem **282**(18): 13532-41.

Low, D. A., N. J. Weyand, et al. (2001). "Roles of DNA adenine methylation in regulating bacterial gene expression and virulence." Infect Immun **69**(12): 7197-204.

Lyon, M. F. (1961). "Gene action in the X-chromosome of the mouse (Mus musculus L.)." Nature **190**: 372-3.

Macleod, D., R. R. Ali, et al. (1998). "An alternative promoter in the mouse major histocompatibility complex class II I-Abeta gene: implications for the origin of CpG islands." Mol Cell Biol **18**(8): 4433-43.

Macleod, D., J. Charlton, et al. (1994). "Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island." Genes Dev **8**(19): 2282-92.

Mardis, E. R. (2007). "ChIP-seq: welcome to the new frontier." Nat Methods **4**(8): 613-4.

Matangkasombut, O. and S. Buratowski (2003). "Different sensitivities of bromodomain factors 1 and 2 to histone H4 acetylation." Mol Cell **11**(2): 353-63.

Mayer, W., A. Niveleau, et al. (2000). "Demethylation of the zygotic paternal genome." Nature **403**(6769): 501-2.

McGrath, J. and D. Solter (1984). "Completion of mouse embryogenesis requires both the maternal and paternal genomes." Cell **37**(1): 179-83.

McKeon, C., H. Ohkubo, et al. (1982). "Unusual methylation pattern of the alpha 2 (I) collagen gene." Cell **29**(1): 203-10.

Meehan, R. R., J. D. Lewis, et al. (1989). "Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs." Cell **58**(3): 499-507.

Meissner, A., T. S. Mikkelsen, et al. (2008). "Genome-scale DNA methylation maps of pluripotent and differentiated cells." Nature **454**(7205): 766-70.

Meselson, M., R. Yuan, et al. (1972). "Restriction and modification of DNA." Annu Rev Biochem **41**: 447-66.

Messmer, S., A. Franke, et al. (1992). "Analysis of the functional role of the Polycomb chromo domain in Drosophila melanogaster." Genes Dev **6**(7): 1241-54.

Metivier, R., R. Gallais, et al. (2008). "Cyclical DNA methylation of a transcriptionally active promoter." Nature **452**(7183): 45-50.

Mikkelsen, T. S., J. Hanna, et al. (2008). "Dissecting direct reprogramming through integrative genomic analysis." Nature **454**(7200): 49-55.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-60.

Miller, T., N. J. Krogan, et al. (2001). "COMPASS: a complex of proteins associated with a trithorax-related SET domain protein." Proc Natl Acad Sci U S A **98**(23): 12902-7.

Milne, T. A., S. D. Briggs, et al. (2002). "MLL targets SET domain methyltransferase activity to Hox gene promoters." Mol Cell **10**(5): 1107-17.

Mizuguchi, G., T. Tsukiyama, et al. (1997). "Role of nucleosome remodeling factor NURF in transcriptional activation of chromatin." Mol Cell **1**(1): 141-50.

Mohandas, T., R. S. Sparkes, et al. (1981). "Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation." Science **211**(4480): 393-6.

Murton, B. L., W. L. Chin, et al. (2010). "Characterising the binding specificities of the subunits associated with the KMT2/Set1 histone lysine methyltransferase." J Mol Biol **398**(4): 481-8.

Nan, X. and A. Bird (2001). "The biological functions of the methyl-CpG-binding protein MeCP2 and its implication in Rett syndrome." Brain Dev **23 Suppl 1**: S32-7.

Nan, X., H. H. Ng, et al. (1998). "Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex." Nature **393**(6683): 386-9.

Ng, H. H., F. Robert, et al. (2003). "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." Mol Cell **11**(3): 709-19.

Ng, K., D. Pullirsch, et al. (2007). "Xist and the order of silencing." EMBO Rep **8**(1): 34-9.

Norris, D. P., N. Brockdorff, et al. (1991). "Methylation status of CpG-rich islands on active and inactive mouse X chromosomes." Mamm Genome **1**(2): 78-83.

Norris, D. P., D. Patel, et al. (1994). "Evidence that random and imprinted Xist expression is controlled by preemptive methylation." Cell **77**(1): 41-51.

Noyer-Weidner, M. and T. A. Trautner (1993). "Methylation of DNA in prokaryotes." EXS **64**: 39-108.

Okano, M., S. Xie, et al. (1998). "Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases." Nat Genet **19**(3): 219-20.

Okano, M., S. Xie, et al. (1998). "Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells." Nucleic Acids Res **26**(11): 2536-40.

Ooi, S. K. and T. H. Bestor (2008). "Cytosine methylation: remaining faithful." Curr Biol **18**(4): R174-6.

Ooi, S. K., C. Qiu, et al. (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." Nature **448**(7154): 714-7.

Oswald, J., S. Engemann, et al. (2000). "Active demethylation of the paternal genome in the mouse zygote." Curr Biol **10**(8): 475-8.

Ozsolak, F., J. S. Song, et al. (2007). "High-throughput mapping of the chromatin structure of human promoters." Nat Biotechnol **25**(2): 244-8.

Park, J. G. and V. M. Chapman (1994). "CpG island promoter region methylation patterns of the inactive-X-chromosome hypoxanthine phosphoribosyltransferase (Hprt) gene." Mol Cell Biol **14**(12): 7975-83.

Pluta, A. F., A. M. Mackay, et al. (1995). "The centromere: hub of chromosomal activities." Science **270**(5242): 1591-4.

Pokholok, D. K., C. T. Harbison, et al. (2005). "Genome-wide map of nucleosome acetylation and methylation in yeast." Cell **122**(4): 517-27.

Pradhan, M., P. O. Esteve, et al. (2008). "CXXC domain of human DNMT1 is essential for enzymatic activity." Biochemistry **47**(38): 10000-9.

Proffitt, J. H., J. R. Davie, et al. (1984). "5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA." Mol Cell Biol **4**(5): 985-8.

Ramirez-Carrozzi, V. R., D. Braas, et al. (2009). "A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling." Cell **138**(1): 114-28.

Reese, B. E., K. E. Bachman, et al. (2003). "The methyl-CpG binding protein MBD1 interacts with the p150 subunit of chromatin assembly factor 1." Mol Cell Biol **23**(9): 3226-36.

Reik, W. (2007). "Stability and flexibility of epigenetic gene regulation in mammalian development." Nature **447**(7143): 425-32.

Reik, W., W. Dean, et al. (2001). "Epigenetic reprogramming in mammalian development." Science **293**(5532): 1089-93.

Robertson, G., M. Hirst, et al. (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." Nat Methods **4**(8): 651-7.

Roh, T. Y., S. Cuddapah, et al. (2006). "The genomic landscape of histone modifications in human T cells." Proc Natl Acad Sci U S A **103**(43): 15782-7.

Rougier, N., D. Bourc'his, et al. (1998). "Chromosome methylation patterns during mammalian preimplantation development." Genes Dev **12**(14): 2108-13.

Russell, M., P. Berardi, et al. (2006). "Grow-ING, Age-ING and Die-ING: ING proteins link cancer, senescence and apoptosis." Exp Cell Res **312**(7): 951-61.

Ruthenburg, A. J., C. D. Allis, et al. (2007). "Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark." Mol Cell **25**(1): 15-30.

Santos-Rosa, H., R. Schneider, et al. (2002). "Active genes are tri-methylated at K4 of histone H3." Nature **419**(6905): 407-11.

Sarraf, S. A. and I. Stancheva (2004). "Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly." Mol Cell **15**(4): 595-605.

Sasaki, H., K. Ishihara, et al. (2000). "Mechanisms of Igf2/H19 imprinting: DNA methylation, chromatin and long-distance gene regulation." J Biochem **127**(5): 711-5.

Sauvageau, M. and G. Sauvageau (2008). "Polycomb group genes: keeping stem cell activity in balance." PLoS Biol **6**(4): e113.

Saxonov, S., P. Berg, et al. (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." Proc Natl Acad Sci U S A **103**(5): 1412-7.

Schaefer, M. and F. Lyko (2010). "Solving the Dnmt2 enigma." Chromosoma **119**(1): 35-40.

Schanen, N. C. (2006). "Epigenetics of autism spectrum disorders." Hum Mol Genet **15 Spec No 2**: R138-50.

Schneider, J., A. Wood, et al. (2005). "Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression." Mol Cell **19**(6): 849-56.

Schones, D. E., K. Cui, et al. (2008). "Dynamic regulation of nucleosome positioning in the human genome." Cell **132**(5): 887-98.

Schubeler, D., D. M. MacAlpine, et al. (2004). "The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote." Genes Dev **18**(11): 1263-71.

Selker, E. U., N. A. Tountas, et al. (2003). "The methylated component of the Neurospora crassa genome." Nature **422**(6934): 893-7.

Shen, L., Y. Kondo, et al. (2007). "Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters." PLoS Genet **3**(10): 2023-36.

Shilatifard, A. (2008). "Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation." Curr Opin Cell Biol **20**(3): 341-8.

Shogren-Knaak, M., H. Ishii, et al. (2006). "Histone H4-K16 acetylation controls chromatin structure and protein interactions." Science **311**(5762): 844-7.

Singh, H. (2009). "Teeing up transcription on CpG islands." Cell **138**(1): 14-6.

Skene, P. J., R. S. Illingworth, et al. (2010). "Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state." Mol Cell **37**(4): 457-68.

Stein, R., A. Razin, et al. (1982). "In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells." Proc Natl Acad Sci U S A **79**(11): 3418-22.

Stirzaker, C., D. S. Millar, et al. (1997). "Extensive DNA methylation spanning the Rb promoter in retinoblastoma tumors." Cancer Res **57**(11): 2229-37.

Straussman, R., D. Nejman, et al. (2009). "Developmental programming of CpG island methylation profiles in the human genome." Nat Struct Mol Biol.

Surani, M. A., S. C. Barton, et al. (1984). "Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis." Nature **308**(5959): 548-50.

Suzuki, M., S. Sato, et al. (2007). "A new class of tissue-specifically methylated regions involving entire CpG islands in the mouse." Genes Cells **12**(12): 1305-14.

Suzuki, M. M., A. R. Kerr, et al. (2007). "CpG methylation is targeted to transcription units in an invertebrate genome." Genome Res **17**(5): 625-31.

Tahiliani, M., K. P. Koh, et al. (2009). "Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by the MLL Fusion Partner TET1." Science.

Takahashi, Y. H., J. S. Lee, et al. (2009). "Regulation of H3K4 Trimethylation via Cps40 (Spp1) of COMPASS is monoubiquitination independent: implication for a Phe/Tyr switch by the catalytic domain of Set1." Mol Cell Biol **29**(13): 3478-86.

Takai, D. and P. A. Jones (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proc Natl Acad Sci U S A **99**(6): 3740-5.

Tate, C. M., M. L. Fishel, et al. (2009). "Embryonic stem cells lacking the epigenetic regulator Cfp1 are hypersensitive to DNA-damaging agents and

exhibit decreased Ape1/Ref-1 protein expression and endonuclease activity." DNA Repair (Amst) **8**(12): 1411-23.

Tazi, J. and A. Bird (1990). "Alternative chromatin structure at CpG islands." Cell **60**(6): 909-20.

Thomson, J. P., P. J. Skene, et al. (2010). "CpG islands influence chromatin structure via the CpG-binding protein Cfp1." Nature **464**(7291): 1082-6.

Tsukada, Y., J. Fang, et al. (2006). "Histone demethylation by a family of JmjC domain-containing proteins." Nature **439**(7078): 811-6.

Vaquero, A., R. Sternglanz, et al. (2007). "NAD+-dependent deacetylation of H4 lysine 16 by class III HDACs." Oncogene **26**(37): 5505-20.

Vermeulen, M., K. W. Mulder, et al. (2007). "Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4." Cell **131**(1): 58-69.

Vetting, M. W., S. d. C. LP, et al. (2005). "Structure and functions of the GNAT superfamily of acetyltransferases." Arch Biochem Biophys **433**(1): 212-26.

Virmani, A. K., A. Rathi, et al. (2001). "Aberrant methylation of the adenomatous polyposis coli (APC) gene promoter 1A in breast and lung carcinomas." Clin Cancer Res **7**(7): 1998-2004.

Visel, A., M. J. Blow, et al. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature **457**(7231): 854-8.

Voo, K. S., D. L. Carlone, et al. (2000). "Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a

CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1." Mol Cell Biol **20**(6): 2108-21.

Wang, Z., C. Zang, et al. (2009). "Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes." Cell **138**(5): 1019-31.

Weber, M., J. J. Davies, et al. (2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." Nat Genet **37**(8): 853-62.

Weber, M. and D. Schubeler (2007). "Genomic patterns of DNA methylation: targets and function of an epigenetic mark." Curr Opin Cell Biol **19**(3): 273-80.

Whiteford, N., T. Skelly, et al. (2009). "Swift: Primary Data Analysis for the Illumina Solexa Sequencing Platform." Bioinformatics.

Wolf, S. F., D. J. Jolly, et al. (1984). "Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation." Proc Natl Acad Sci U S A **81**(9): 2806-10.

Wyatt, G. R. and S. S. Cohen (1952). "A new pyrimidine base from bacteriophage nucleic acids." Nature **170**(4338): 1072-3.

Wysocka, J., M. P. Myers, et al. (2003). "Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1." Genes Dev **17**(7): 896-911.

Wysocka, J., T. Swigut, et al. (2005). "WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development." Cell **121**(6): 859-72.

Xiao, T., Y. Shibata, et al. (2007). "The RNA polymerase II kinase Ctk1 regulates positioning of a 5' histone methylation boundary along genes." Mol Cell Biol **27**(2): 721-31.

Xie, S., Z. Wang, et al. (1999). "Cloning, expression and chromosome locations of the human DNMT3 gene family." Gene **236**(1): 87-95.

Yamada, T., A. R. Carson, et al. (2005). "Endothelial nitric-oxide synthase antisense (NOS3AS) gene encodes an autophagy-related protein (APG9-like2) highly expressed in trophoblast." J Biol Chem **280**(18): 18283-90.

Yang, C., E. Bolotin, et al. (2007). "Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters." Gene **389**(1): 52-65.

Yoder, J. A., N. S. Soman, et al. (1997). "DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe." J Mol Biol **270**(3): 385-95.

Yokoyama, A., Z. Wang, et al. (2004). "Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression." Mol Cell Biol **24**(13): 5639-49.

Yoon, J. H., S. Iwai, et al. (2003). "Human thymine DNA glycosylase (TDG) and methyl-CpG-binding protein 4 (MBD4) excise thymine glycol (Tg) from a Tg:G mispair." Nucleic Acids Res **31**(18): 5399-404.

Young, S. R. and D. G. Skalnik (2007). "CXXC finger protein 1 is required for normal proliferation and differentiation of the PLB-985 myeloid cell line." DNA Cell Biol **26**(2): 80-90.

Yu, B. D., J. L. Hess, et al. (1995). "Altered Hox expression and segmental identity in Mll-mutant mice." Nature **378**(6556): 505-8.

Zhang, X., J. Yazaki, et al. (2006). "Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis." Cell **126**(6): 1189-201.

Zhang, Y., H. H. Ng, et al. (1999). "Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation." Genes Dev **13**(15): 1924-35.

Zhao, X. D., X. Han, et al. (2007). "Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells." Cell Stem Cell **1**(3): 286-98.

Appendix A

1. Thomson, J. P., P. J. Skene, et al. "CpG islands influence chromatin structure via the CpG-binding protein Cfp1." Nature 464(7291): 1082-6.